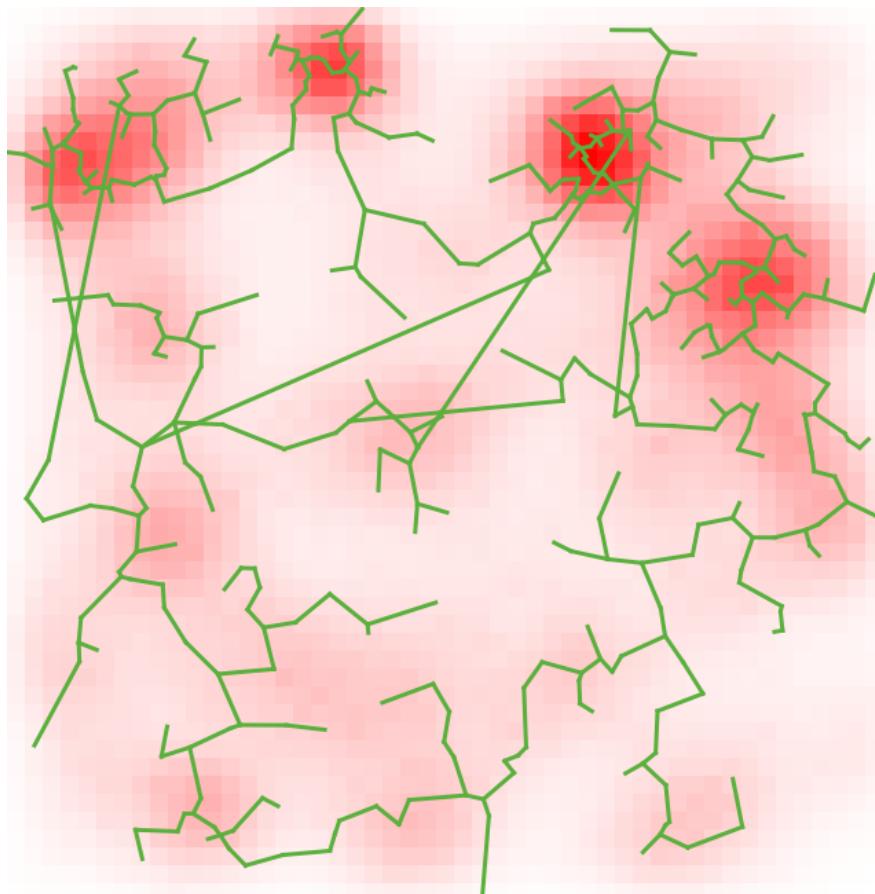


VERS DES MODÈLES COUPLANT DÉVELOPPEMENT URBAIN ET CROISSANCE DES RÉSEAUX DE TRANSPORT

JUSTE RAIMBAULT



Mémoire de Thèse de Doctorat

Under the supervision of ARNAUD BANOS and FLORENT LE NÉCHET

UMR CNRS 8504 Géographie-cités
and UMR-T IFSTTAR 9403 LVMT

Université Paris Diderot - Paris 7

June 2017 – version 3.1

Juste Raimbault : *Vers des Modèles Couplant Développement Urbain et Croissance des Réseaux de Transport*, Mémoire de Thèse de Doctorat, © June 2017

ABSTRACT

Résumé

C : (Florent) trop de concepts dans l'abstract, peut pas apporter qqchse à tous

C : (Florent) commencer par expliquer ce que sont causalités circulaires et pourquoi difficiles à modéliser

C : (Arnaud) complexly :?

C : (Arnaud) théorie des systèmes territoriaux en réseau co-évolutifs ?

READING NOTES

Notes de Lecture

This provisory Memoire must be read as a work in progress, as it details progresses after one year of Doctorate. Many parts are given at the state of project, and not omitted as playing a role in the current research questioning. Its purpose is to set up a plan and examine the achieved work and corresponding directions, but also to share research ideas at this important step of one year.

PUBLICATIONS

Les travaux suivants contiennent une grande partie du contenu de cette thèse :

PUBLICATIONS

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016.*

Raimbault, J. (2017). A Discrepancy-Based Framework to Compare Robustness Between Multi-attribute Evaluations. In *Complex Systems Design & Management* (pp. 141-154). Springer International Publishing. [RAIMBAULT, 2016b]

Raimbault, J. (2016). Investigating the Empirical Existence of Static User Equilibrium, *forthcoming in EWGT 2016 proceedings, Transportation Research Procedia*. arxiv :1608.05266 [RAIMBAULT, 2016e]

Raimbault, J. (2016). Generation of Correlated Synthetic Data, forthcoming in *Actes des Journées de Rochebrune 2016*.

Raimbault, J. (2015). Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach, forthcoming in *ECTQG 2015 proceedings*. arxiv :1605.08888

COMMUNICATIONS

Towards a Theory of Co-evolutive Networked Territorial Systems : Insights from Transportation Governance Modeling in Pearl River Delta, China, *MEDIUM Seminar : Sustainable Development in Zhuhai, Guangzhou, Dec 2016*.

Models of growth for system of cities : Back to the simple, *Conference on Complex Systems 2016, Amsterdam, Sep 2016*.

For a Cautious Use of Big Data and Computation. *Royal Geographical Society - Annual Conference 2016 - Session : Geocomputation, the Next 20 Years (1), London, Aug 2016*.

Indirect Bibliometrics by Complex Network Analysis. *20e Anniversaire de Cybergeo, Paris, May 2016*.

Raimbault, J. & Serra, H. (2016). Game-based Tools as Media to Transmit Freshwater Ecology Concepts, *poster corner at SETAC 2016 (Nantes, May 2016)*.

Le Néchet, F. & Raimbault, J. (2015). Modeling the emergence of metropolitan transport authority in a polycentric urban region, *ECTQG 2015, Bari, Sep 2015*.

Hybrid Modeling of a Bike-Sharing Transportation System, *poster presented at ICCSS 2015, Helsinki, June 2015*.

Raimbault, J. & Gonzales, J. (2015). Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport, *poster presented at Rencontres du Labex Dynamite, Paris, May 2015*.

ACKNOWLEDGEMENTS

Les résultats obtenus dans la section 3.1 de cet article ont été calculés sur l’organisation virtuelle vo.complex-system.eu de l’European Grid Infrastructure (<http://www.egi.eu>). Nous remercions l’European Grid Infrastructure et ses National Grid Initiatives (France-Grilles en particulier) pour fournir le support technique et l’infrastructure.

TABLE DES MATIÈRES

Introduction	3
I FOUNDATIONS	17
1 INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	19
1.1 Réseaux et Territoires	21
1.2 De Paris à Zhuhai	30
1.3 Elements de terrain	31
2 MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	37
2.1 Modéliser les Interactions	38
2.2 Une Approche Epistémologique	47
2.3 Revue Systématique et Modélographie	55
3 POSITIONNEMENTS	59
3.1 Reproductibilité	61
3.2 Calcul Intensif et Exploration des Modèles	70
3.3 Positionnement Epistémologique	83
II BRIQUES ÉLÉMENTAIRES	91
4 THÉORIE EVOLUTIVE URBAINE	93
4.1 Corrélations Statiques	94
4.2 Causalités Spatio-temporelles	100
4.3 Effets de Réseaux	111
5 ECHELLES ET ONTOLOGIES	113
5.1 Equilibre Utilisateur Statique	115
5.2 Transport Routier et Déterminants des Coûts	128
5.3 Transactions immobilières et Grand Paris	141
6 MORPHOGENÈSE URBAINE	147
6.1 Une Approche Interdisciplinaire de la Morphogenèse .	148
6.2 Morphogenèse Urbaine par Agrégation-diffusion . .	150
6.3 Génération de configurations territoriales corrélées .	151
III SYNTHESIS : CONSTRUCTION OF CO-EVOLUTION MODELS	163
7 CO-ÉVOLUTION À L'ECHELLE MACROSCOPIQUE	169
7.1 Exploration de SimpopNet	170
7.2 Modèle d'interaction	171
7.3 Le Modèle SimpopSino	173
8 CO-EVOLUTION AT THE MESO-SCALE	175
8.1 Modèles de Croissance de Réseau	176
8.2 Co-évolution à l'échelle mesoscopique	177
8.3 Gouvernance du Système de Transport	178
9 CADRE THÉORIQUE	181

9.1	Une Théorie Géographique	183
9.2	Un Cadre pour les Systèmes Socio-techniques	191
9.3	Un Cadre de Connaissances Appliqué	203
	Conclusion	221
	BIBLIOGRAPHIE	231
	IV APPENDICES	261
A	INFORMATIONS SUPPLÉMENTAIRES	263
A.1	Elements de Terrain	263
A.2	Technical Developments	263
A.3	Exploration des Modèles	267
A.4	Causalités dans le modèle RBD	268
B	METHODOLOGICAL DEVELOPMENTS	269
B.1	Un cadre uniifié pour les modèles stochastiques de croissance urbaine	270
B.2	Sensibilité des Lois d’Echelle Urbaines à l’Etendue Spatiale	273
B.3	Correlations spatio-temporelles	275
B.4	Génération de Données Synthétiques Corrélées	278
B.5	Un Cadre basé sur la Discrépance	281
C	DÉVELOPPEMENTS THÉMATIQUES	297
C.1	An Interdisciplinary Approach to Morphogenesis	298
C.2	Generation of Correlated Synthetic Data	299
C.3	Classifying Patents Based on their Semantic Content	305
D	DONNÉES	329
D.1	Données de Traffic du Grand Paris	329
D.2	Prix de l’Essence aux Etats-Unis	329
D.3	Réseau Routier Européen	329
D.4	Réseau Dynamique des Autoroutes Françaises	329
D.5	Interviews	329
E	OUTILS	331
E.1	Softwares and Packages	332
E.2	Architecture and Sources for Algorithms and Models of Simulation	333
E.3	Tools and Workflow for an open Reproducible Research	337
F	QUANTITATIVE ANALYSIS OF THESIS REFLEXIVITY	339

TABLE DES FIGURES

FIGURE 1	Architecture de l'algorithme de revue systématique	50
FIGURE 2	-NoValue-	51
FIGURE 3	Représentation schématique de la distinction entre différents types de modèles couplant territoires et réseaux.	56
FIGURE 4	Reproductibilité et visualisation	64
FIGURE 5	Usage naïf de la fouille de données et du calcul intensif	74
FIGURE 6	Distance des diagramme de phase à la référence	80
FIGURE 7	Exemples de diagrammes de phase	80
FIGURE 8	Distribution spatiale des indicateurs morphologiques	95
FIGURE 9	-NoValue-	95
FIGURE 10	Correlations dans le modèle RDB	106
FIGURE 11	Identification de régimes d'interactions	107
FIGURE 12	Application web pour les données de trafic	119
FIGURE 13	Variabilité spatiale des plus courts chemins	120
FIGURE 14	Variabilité des temps et distance de trajet	121
FIGURE 15	Stabilité temporelle de la centralité de chemin maximale	122
FIGURE 16	Auto-corrélation spatiale pour les temps de trajet relatifs	124
FIGURE 17	Prix moyen par Contés	132
FIGURE 18	Index d'autocorrelation spatiale de Moran	134
FIGURE 19	Résultats des analyses GWR	137
FIGURE 20	Projets de transport successifs du Grand Paris	143
FIGURE 21	Corrélations retardées empiriques	144
FIGURE 22	Exploration de l'espace faisable des corrélations	156
FIGURE 23	Exemple de génération de configurations couplées	157
FIGURE 24	Schématisation du modèle	171
FIGURE 25	Croissance de réseau biologique	176
FIGURE 26	Réseau de citations de la Théorie Evolutive Urbaine	209
FIGURE 27	Réseau complet des domaines de connaissance	213
FIGURE 28	-NoValue-	267
FIGURE 29	268
FIGURE 30	Cartes de ségrégation métropolitaine	291
FIGURE 31	Sensibilité de la robustesse aux données manquantes	293

FIGURE 32	-NoValue-	301
FIGURE 33	-NoValue-	303
FIGURE 34	-NoValue-	304
FIGURE 35	314
FIGURE 36	315
FIGURE 37	315
FIGURE 38	317
FIGURE 39	320
FIGURE 40	320
FIGURE 41	322
FIGURE 42	322
FIGURE 43	324

LISTE DES TABLEAUX

TABLE 1	Proximités lexicales stationnaires	52
TABLE 2	Statistiques descriptives des prix des carburants (\$ par gallon)	131
TABLE 3	Régressions au niveau du Conté	139
TABLE 4	165
TABLE 5	Résultats numériques des simulations synthétiques	290
TABLE 6	328

C : (Florent) cf receuil articles du Monde sur Grd Paris (numériser)

C : (Florent) HDR Anne ?

C : (Florent) trop peu ancré concrètement dans le champ des interactions transport/ville - enchainement idée ok mais revoir granularité info. Catalogue de situations complexes d'interactions forme urbaine/transport à reproduire.

C (Arnaud) : titre : répétitif sur "growth"; P7 : Paris 7 Paris Diderot

A1 : (JR) ok titre en français

C : (Arnaud) Sur le plan : Contexte; Théorique; cadre Methodo. Quant epistemo avant methodo ou théorie? Publier Cybergeo; Article ectqg en annexe; dans méthodo : ajouter Modèle agents - hors équilibre; reorg Empirical etc (Q/T)

C : (Arnaud) A LIRE : R Brunet, Discontinuités en géographie; Pierre Dumolard (Espace Différencié); Guy DiMeo (L'homme, la société, l'espace)

INTRODUCTION

INTRODUCTION

Introduction

C'est quand on donne un coup de pied dans la fourmilière qu'on se rend compte de toute sa complexité.

- ARNAUD BANOS

C : (Florent) cet exemple paraît loin de l'approche ? il n'est pas territorial, bien mais HS ; un autre sur dynamiques de certaines villes connectées ou non serait plus approprié

"En conséquence d'un problème technique, le trafic est interrompu sur la ligne B du RER pour une durée indéterminée. Plus d'information seront fournies dès que possible". Il y a des fortes chances pour que quiconque ayant vécu ou passé un peu de temps en région parisienne ait déjà entendu cette annonce glaçante et en ait subi les conséquences pour le reste de la journée. Mais il ne se doute sûrement pas des ramifications des cascades causales induites par cet évènement presque banal. Les systèmes territoriaux, quelles que soient les aspects considérés pour leur définition, seront toujours extrêmement complexes, les interrelations à de nombreuses échelles spatiales et temporelles participant à la production des comportements émergents observés à tout niveau du système. Martin est un étudiant qui fait l'aller-retour journalier entre Paris et Palaiseau and manquera un examen crucial, ce qui aura un impact profond sur sa vie professionnelle : implications à une longue échelle de temps, une petite échelle spatiale et à la granularité de l'agent. **C : (Florent) ?** Yuangsi était en train de relier les aéroports d'Orly et Roissy dans son voyage de Londres à Pékin et va manquer son avion ainsi que le mariage de sa soeur : grande échelle spatiale, petite échelle de temps, granularité de l'agent. Une pétition collective émerge des voyageurs, conduisant à la création d'une organisation qui mettra la pression sur les autorités pour qu'elles augmentent le niveau de service : échelle temporelle et spatiales mesoscopique, granularité de l'aggregation d'agents. La recherche de cause possible à l'incident conduira à des processus intriqués à diverses échelles, parmi lesquels aucun ne semble être une meilleure explication ; le développement historique du réseau ferroviaire en région parisienne a conditionné les évolutions futures et le RER B a suivi l'ancienne Ligne de Sceaux, le plan de DELOUVRIER pour le développement régional et son execution partielle, sont également des éléments d'explication des faiblesses structurelles du réseau parisien de transports en commun [GLEYZE, 2005] **C : (Florent) réseau parisien un des plus résilients du monde, cf slides Erik Janus**

KTH ; les motifs pendulaires dus à l'organisation territoriale induisent une surcharge de certaines ligne et ainsi nécessairement une augmentation des incidents d'exploitation. La liste pourrait être ainsi continuée un certain temps, chaque approche apportant sa vision mature correspondant à un corpus de connaissances scientifiques dans des disciplines diverses comme la géographie, l'économie urbains, les transports. Cette anecdote amusante est suffisante pour faire ressentir la complexité des systèmes territoriaux. Notre but ici est de se plonger dans cette complexité, et en particulier donner un point de vue original sur l'étude des relations entre réseaux et territoires. Le choix de cette position sera largement discuté dans une partie thématique, nous nous concentrerons à présent sur l'originalité du point de vue que nous allons prendre.

DE LA POSITION GÉNÉRALE

L'ambition de cette thèse est de ne pas avoir d'ambition. Cette entrée en matière, rude en apparence, contient à différents niveaux les logiques sous-jacentes à notre processus de recherche. Au sens propre, nous nous plaçons tant que possible dans une démarche constructive et exploratoire, autant sur les plans théoriques et méthodologiques que thématique, mais encore proto-méthodologique (outils appliquant la méthode) : si des ambitions unidimensionnelles ou intégrées devaient émerger, elles seraient conditionnées par l'arbitraire choix d'un échantillon temporel parmi la continuité de la dynamique qui structure tout projet de recherche. Au sens structurel, l'auto-référence qui soulève une contradiction apparente met en exergue l'aspect central de la réflexivité dans notre démarche constructive, autant au sens de la récursivité des appareils théoriques, de celui de l'application des outils et méthodes développés au travail lui-même ou que de celui de la co-construction des différentes approches et des différents axes thématiques. Le processus de production de connaissance pourra ainsi être lu comme une métaphore des processus étudiés. Enfin, sur un plan plus enclin à l'interprétation, cela suggérera la volonté d'une position délicate liant un positionnement politique dont la nécessité est intrinsèque aux sciences humaines (par exemple ici contre l'application technocratique des modèles, ou pour le développement d'outils luttant pour une science ouverte) à une rigueur d'objectivité plus propre aux autres champs abordés, position forçant à une prudence accrue.

CONTEXTE SCIENTIFIQUE : PARADIGMES DE LA COMPLEXITÉ

Pour une meilleure introduction du sujet, il est nécessaire d'insister sur le cadre scientifique dans lequel nous nous positionnons. Ce contexte est crucial à la fois pour comprendre les concepts épistémo-

logiques implicites dans nos questions de recherche, et aussi pour être conscient de la variété de méthodes et outils utilisés. La science contemporaine prend progressivement le tournant de la complexité dans de nombreux champs **C : (Florent) tout le monde ne connaît pas**, ce qui implique une mutation épistémologique pour abandonner le réductionnisme strict qui a échoué dans la majorité de ses tentatives de synthèse [ANDERSON, 1972]. Arthur a rappelé récemment [ARTHUR, 2015] qu'une mutation des méthodes et paradigmes en était également un enjeu, de par la place grandissante prise par les approches computationnelles qui remplacent les résolutions purement analytiques généralement limité en possibilités de modélisation et de résolution. La capture des *propriétés émergentes* par des modèles de systèmes complexes est une des façons d'interpréter la philosophie de ces approches.

C : (Florent) rebondir sur thématique, questce qui emerge

Ces considérations sont bien connues des Sciences Humaines (qualitatives et quantitatives) pour lesquelles la complexité des agents et systèmes étudiés est une des justifications de leur existence : si les humains étaient des particules, la majorité des disciplines les prenant comme objet d'étude n'auraient jamais émergé puisque la thermodynamique aurait alors résolu la majorité des problèmes sociaux **C : (Florent)attention phrases asimoviennes**¹. Elles sont au contraire moins connues et acceptées en sciences "dures" comme la physique : LAUGHLIN développe dans [LAUGHLIN, 2006] une vision de la discipline **C : (Florent)which?** à la même position de "frontière des connaissances" que d'autre champs pouvant paraître moins matures. La plupart des connaissances actuelles concerne des structures classiques simples, alors qu'un grand nombre de système présentent des propriétés *d'auto-organisation*, au sens où les lois macroscopiques ne sont pas suffisantes pour inférer les propriétés macroscopiques du systèmes à moins que son évolution soit entièrement simulée (plus précisément cette vision peut être prise comme une définition de l'émergence sur laquelle nous reviendrons par la suite, or des propriétés auto-organisées sont par nature émergentes). Cela correspond au premier cauchemar du Démon de Laplace développé dans [DEFUANT et al., 2015].

A la croisée de positionnements épistémologiques, de méthodes et de champs d'application, les *Sciences de la complexité* se concentrent sur l'importance de l'émergence et de l'auto-organisation dans la plupart des phénomènes réel, ce qui les place plus proche de la frontière des connaissances que ce que l'on peut penser pour des disciplines classiques (LAUGHLIN, op. cit.). Ces concepts ne sont pas récents et avaient déjà été mis en valeur par ANDERSON [ANDERSON,

¹ bien que cette affirmation soit elle-même discutable, les sciences physiques classiques ayant également échoué à prendre en compte l'irréversibilité et l'évolution de Systèmes Complexes Adaptatifs comme le souligne PRIGOGINE dans [PRIGOGINE et STENGERS, 1997].

1972]. On peut aussi interpréter la Cybernétique comme un précurseur des Sciences de la Complexité en la lisant comme un pont entre technologie et sciences cognitives [WIENER, 1948]. **C : (Florent) pquoiparler de ca ici ?** Plus tard, la Synergétique [HAKEN, 1980] a posé les bases d'approches théoriques des phénomènes collectifs en physique. Les causes possibles de la croissance récente du nombre de travaux se réclamant d'approches complexes sont nombreuses. L'explosion de la puissance de calcul en est certainement une vu le rôle central que jouent les simulations numériques [VARENNE, 2010b]. Elles peuvent aussi être à chercher auprès de progrès en épistémologie : introduction de la notion de perspectivisme [GIERE, 2010c], réflexions plus fine autour de la nature des modèles [VARENNE et SILBERSTEIN, 2013]². Les potentialités théoriques et empiriques de telles approches jouent nécessairement un rôle dans leur succès³, comme le confirme les domaines très variés d'application (voir [NEWMAN, 2011] pour une revue très générale), comme par exemple la Science de Réseaux [BARABASI, 2002] ; les Neurosciences [Koch et LAURENT, 1999] ; les Sciences Sociales ; la Géographie [MANSON, 2001][PUMAIN, 1997] ; la Finance avec les approches écononophysiques [STANLEY et al., 1999] ; l'Ecologie [GRIMM et al., 2005]. La Feuille de Route des Systèmes Complexes [BOURGINE, CHAVALARIAS et AL., 2009] propose une double lecture des travaux en Complexité : une approche horizontale faisant la connexion entre champs d'étude par des questions transversales sur les fondations théoriques de la complexité et des faits stylisés empiriques communs, et une approche verticale, dans le but de construire des disciplines intégrées et les modèles multi-scalaires hétérogènes correspondants. L'interdisciplinarité est ainsi cruciale pour notre contexte scientifique.

C : (Florent) donner ici exemples dans champ transports/urba

C : (Florent) plus de détails sur les disciplines CS ?

INTERDISCIPLINARITÉ

Il est important d'insister sur le rôle de l'interdisciplinarité dans la position de recherche prise ici. Il s'agit moins d'un travail en Géographie ou en Modélisation de Systèmes Complexes Adaptatifs, pouvant difficilement être vraiment les deux à la fois, mais en *Science des Systèmes Complexes* que nous réclamons discipline propre comme le propose PAUL BOURGINE. **C (Florent) : pas vraiment fondateur de la discipline A1 : non mais du point de vue particulier que nous défendons - théories intégratives roadmap etc. - trouver une ref là dessus ?**

² dans ce cadre, les progrès scientifiques et épistémologiques ne peuvent pas être dissociés et peuvent être vus comme étant en co-évolution

³ même si l'adoption de nouvelles pratiques scientifiques est souvent largement biaisé par l'imitation et le manque d'originalité [DIRK, 1999], ou de façon plus ambiguë, par des stratégies de positionnement puisque le combat pour les fonds est un obstacle croissant à une recherche saine [BOLLEN et al., 2014].

Ce n'est pas sans risques d'être lu avec méfiance voir défiance par les tenants des disciplines classiques, comme des exemples récents de malentendus ou conflits ont récemment illustré [DUPUY et BENGUIGUI, 2015]. Il faut se rappeler l'importance de la spirale vertueuse de BANOS entre disciplinarité et interdisciplinarité [BANOS, 2013]. Celle-ci doit nécessairement impliquer différents agents scientifiques, et il est compliqué pour un agent de se positionner dans les deux branches ; notre fond scientifique ne nous permet pas de nous positionner dans la *disciplinarité géographique* mais bien dans celle des Systèmes Complexes (qui est interdisciplinaire, voir 3.3 pour contourner la contradiction apparente), et notre sensibilité scientifique et épistémologique nous pousse à faire de même.

Le positionnement de BATTY lorsqu'il propose *Une Nouvelle Science des Villes* [BATTY, 2013b] (qu'il présente avec humour comme *La nouvelle science des villes*), se présente comme une intégration des disciplines et méthodes vers une science définie par son objet d'étude, les villes. Its theoretical and epistemological weaknesses (no theoretical constructions of studied geographical objects on the one hand, approximative contextualization of complexity) combined with an overall impression of *pot-pourri* of forgotten works (space syntax, land-use models), unfortunately avoid us to use it as we will use geographical theories (e.g. evolutive urban theory) in an appropriated epistemological complexity context. Yet our reading of this work may be the result of a misunderstanding due to different cultural backgrounds.

C (Arnaud) : j'espère que tu abuses ? :)!! Argument d'autorité A1 : yes, changer positionnement complètement malvenu **C (Florent) :** attention arguments autorité; insister sur difficulté à intégrer paradigmes plutôt que juger précédents **A1 :** idem

L'évolution scientifique des sciences de la complexité, qui est vue par certains comme une révolution [COLANDER, 2003], ou même comme *un nouveau type de science*, pourrait affronter des difficultés intrinsèques dues aux comportements et a-priori des chercheurs en tant qu'être humains. **C : (Florent) idem développer transport/transport/modeling (?)**

Plus précisément, le besoin d'interdisciplinarité qui fait la force des Sciences de la Complexité pourrait devenir une de ses grandes faiblesses, puisque la structure fortement en silo de la science peut avoir des impacts négatifs sur les initiatives impliquant des disciplines variées. Nous n'évoquons pas les problèmes de sur-publication, quantification, competition, qui sont plus liés à des questions de Science Ouverte et de son éthique, tout aussi de grande importance mais d'une autre nature. Cette barrière qui nous hante et que nous pourrions ne pas surmonter, a pour plus évident symptôme des *divergences culturelles disciplinaires*, et les conflits d'opinion en résultant. Ce drame du malentendu scientifique est d'autant plus grave qu'il peut en effet détruire totalement certains progrès en interprétant comme une falsification des travaux qui traitent une ques-

tion toute différente. L'exemple récent d'un travail sur les inégalités liées aux hauts revenus présenté dans [AGHION et al., 2015], et dont les conclusions ont été commentées comme s'opposant aux thèses de Piketty dans [PIKETTY, 2013], est typique de ce schéma. Alors que Piketty se concentre sur la construction de bases de données propres sur le temps long pour les revenus et montre empiriquement une récente accélération des inégalités de revenus, son modèle visant à lier ce fait stylisé avec l'accumulation de capital a été critiqué comme sur-simplifié. D'autre part, Bergeaud *et al.* montrent par un modèle d'économie de l'innovation que *sous certaines hypothèses* les écarts de revenus peuvent être bénéfique à l'innovation et donc à une utilité globale. D'où des conclusions divergentes sur le rôles des capitaux personnels dans une économie. **C : (Florent) hors-sujet, reste ds domaine (?)** Mais des *point de vue* ou *interprétations* différentes ne signifient pas une incompatibilité scientifique, et on pourrait même imaginer rassembler ces deux approches dans un cadre et modèle unifié, produisant des interprétations possiblement similaires et potentiellement encore nouvelles. Une telle approche intégrée aura de grandes chances de contenir plus d'information (selon comment le couplage est opéré) et être une avancée scientifique. Cette expérience de pensée illustre les potentialités et la nécessité de l'interdisciplinarité. Dans une autre veine assez similaire, [HOLMES et al., 2017] ré-analyse des données biologiques d'une expérience de 1943 qui prétendait confirmer l'hypothèse des processus d'évolution Darwiniens par rapport aux processus Lamarckiens, et montrent que les conclusions ne tiennent plus dans le contexte actuel d'analyse de données (avances énormes sur la théorie et les possibilités de traitement) et scientifique (avec d'autre nombreuses preuves de nos jours des processus Darwiniens) : c'est un bon exemple de malentendu sur le contexte, et comment le cadre de travail à la fois technique et thématique influence fortement les conclusions scientifiques. Nous développons à présent divers exemples révélateurs de la manière dont des conflits entre disciplines peuvent être dommageables.

LA TENTATION DE RÉINVENTER LA GÉOGRAPHIE Comme déjà mentionné, DUPUY et BENGUIGUI soulignent dans [DUPUY et BENGUIGUI, 2015] le fait que les sciences urbaines **C : (Florent) définition ?** ont récemment connu des conflits ouverts entre les tenants classiques des disciplines et des nouveaux arrivants, en particulier les physiciens. **C : (Florent) gravité de Wilson par max entropie n'est pas nouveau** La disponibilité de grand jeux de données d'un nouveau type (réseaux sociaux, données des nouvelles technologies de la communication) ont attiré leur attention sur des objets plus traditionnellement étudiés par les sciences humaines, puisque les méthodes analytiques et computationnelles de la physique statistique sont devenues applicables. Bien que ces travaux soient généralement présentés

comme la construction d'une approche scientifique des villes, tout en impliquant que la connaissance existante n'est pas scientifique de par sa nature plus qualitative, ils n'ont aucunement révélé de connaissance nouvelle sur les systèmes urbains : **C : (Florent) pas nécessaire dans la thèse** pour citer quelques exemples, [BARTHELEMY et al., 2013] conclut que Paris a subit une transition pendant la période d'Haussman et ses opérations de planification globale, qui sont des faits naturellement connus depuis longtemps en Histoire Urbaine et Géographie Urbaine. [CHEN, 2009] redécouvre que le modèle gravitaire est amélioré par l'introduction de décalages dans les interactions et dérive analytiquement l'expression d'une force d'interaction entre les villes, sans aucun cadre théorique ni thématique. De tels exemples peuvent être multipliés, confirmant l'inconfort courant entre physiciens et géographes. Des bénéfices significatifs pourraient résulter d'une intégration raisonnée des disciplines [O'SULLIVAN et MANSON, 2015] mais la route semble être bien longue encore.

C : (Florent) a développer, concrètement, quels verrous à faire sauter ?

ECONOMIE GÉOGRAPHIE OU GÉOGRAPHIE ECONOMIQUE ? Des conflits similaires se rencontrent en économie : comme décrit par [MARCHIONNI, 2004], la discipline de l'économie géographique, traditionnellement proche de la géographie, a fortement critiqué un nouveau courant de pensé nommé *économie géographisée*, **C : (Arnaud) New economic geography?** dont le but est la spatialization des techniques économiques classiques. Chacune n'ont pas les mêmes desseins et buts, et le conflit apparaît comme un malentendu complet vu d'un oeil extérieur.

C : (Florent) a développer ou ne pas en parler, un peu loin du cœur du sujet tel que abordé

MODÉLISATION BASÉE AGENT EN ECONOMIE Des conflits disciplinaires peuvent aussi se manifester sous la forme d'un rejet de méthodes nouvelles par les courants dominants. Suivant FARMER [FARMER et FOLEY, 2009], l'échec opérationnel de la plupart des approches économiques classiques pourrait être compensé par un usage plus systématique de la modélisation et simulation basées agent. L'absence de cadre analytique qui est naturelle pour l'étude de la plupart des systèmes complexes adaptatifs semble rebuter la plupart des économistes. **C : (Florent) contraire sans doute vrai aussi**

C : (Arnaud) Difficile de se positionner de manière crédible sur ces sujets en 5 lignes et 1 référence !

FINANCE La finance quantitative peut être instructive pour notre propos et sujet, d'une part par les similarités de la cuisine interdisciplinaire avec notre domaine (rapport avec la physique et l'éco-

nomie, champs plus ou moins "rigoureux", etc.). Dans ce domaine coexistent divers champs de recherche ayant très peu d'interactions entre eux. On peut considérer deux exemples. D'une part, les statistiques et l'économétrie sont extrêmement avancées en mathématiques théoriques, utilisant par exemple des méthodes de calcul stochastique et de théorie des probabilités pour obtenir des estimateurs très raffinés de paramètres pour un modèle donné (voir par exemple [BARNDORFF-NIELSEN et al., 2011]). D'autre part, l'éconophysique a pour but d'étudier des faits stylisés empiriques et inférer les lois correspondantes pour tenter d'expliquer les phénomènes liés à la complexité des marchés financiers [STANLEY et al., 1999], comme par exemple les cascades menant aux ruptures de marché, les propriétés fractales des signaux des actifs, la structure complexe des réseaux de corrélation. Chacun a ses avantages dans un contexte particulier et gagnerait à des interactions accrues entre les deux domaines.

Ces divers exemples pris au fil du vent sont de brèves illustrations du caractère crucial de l'interdisciplinarité et de sa difficulté à pratiquer. Sans presque exagérer, on pourrait imaginer l'ensemble des chercheurs se plaindre de mauvaises ou difficiles expériences d'interdisciplinarité, avec un retour largement positif lors des rares succès. Nous allons tenter par la suite d'emprunter ce chemin étroit, empruntant des idées, théories et méthodes de diverse disciplines, dans l'idéal de la construction d'une connaissance intégrée. En effet, le couplage d'approches hétérogènes à différents niveaux et échelles **C : (Florent) différence ?** sera une clé de voute de cette thèse, la moelle épinière de la philosophie sous-jacente et une composante de la théorie qu'on construira.

C : (Florent) non, disent que difficultés existent mais pas lesquelles, et surtout pas dans le champ d'investigation à venir

TODO : également un développement sur "quanti-quali"

PARADIGMES DE LA COMPLEXITÉ EN GÉOGRAPHIE

Pour revenir à notre anecdote introductory, nous nous concentrerons sur l'étude d'un objet thématique qui sera les systèmes territoriaux : à l'échelle microscopique, les agents peuvent bien être vus comme éléments constitutifs fondamentaux du territoire, qui émergera comme processus complexe à différentes échelles. Plus généralement, il s'agit par commencer de brosser une revue du rôle de la complexité en géographie. Les géographes sont familiers avec la complexité depuis un certain temps, puisque l'étude des interactions spatiales est l'un de ses objets de prédilection. La variété de champs en géographie (géomorphologie, géographie physique, géographie environnementale, géographie humaine, géographie de la santé, etc. pour en nommer quelques) a sûrement joué un rôle clé dans la constitution d'une

pensée géographique subtile, qui considère des processus hétérogènes et multi-scalaires.

PUMAIN rappelle dans [PUMAIN, 2003] une histoire subjective de l'émergence des paradigmes de la complexité en géographie. La cybernétique a produit des théories des systèmes comme celle utilisée par Forrester. **C : (Florent) pas dvlpé, difficile à lire** Plus tard, le glissement vers les concepts de criticalité auto-organisée et d'auto-organisation en physique ont conduit aux développements correspondants en géographie, comme [SANDERS, 1992] qui témoigne de l'application des concepts de la synergétique aux dynamiques des systèmes urbains. Enfin, les paradigmes actuels des systèmes complexes ont été introduits par plusieurs entrées. Par exemple, la nature fractale de la forme urbaine a été introduite par [BATTY et LONGLEY, 1994] et a eu de nombreuses applications jusqu'à des développements plus récents [KEERSMAECKER, FRANKHAUSER et THOMAS, 2003]. BATTY a aussi introduit les automates cellulaires en modélisation urbaine et propose une synthèse jointe avec les modèles basés agents et les fractales dans [BATTY, 2007]. Une autre introduction de la complexité en géographie fut pour le cas des systèmes urbains à travers la théorie évolutive des villes de PUMAIN. En interaction intime avec la modélisation dès ses débuts (le premier modèle Simpop décrit par [SANDERS et al., 1997] rentre dans le cadre théorique de [PUMAIN, 1997]), cette théorie vise à comprendre les systèmes de villes comme des systèmes d'agents adaptatifs en co-évolution, aux interactions multiples, avec différents aspects mis en valeur comme l'importance de la diffusion des innovations. La série des modèles Simpop [PUMAIN, 2012a] a été conçue pour tester différentes hypothèses de la théorie, comme par exemple le rôle des processus de diffusion de l'innovation dans l'organisation du système urbain. Ainsi, des régimes sous-jacent différents ont été mis en évidence pour les systèmes de ville en Europe et aux Etats-unis [BRETAGNOLLE et PUMAIN, 2010a]. A d'autres échelles de temps et dans d'autres contextes, le modèle SimpopLocal [SCHMITT, 2014] a pour but d'étudier les conditions pour l'émergence de systèmes urbains hiérarchiques à partir d'établissements disparates. Un modèle minimal (au sens de paramètres nécessaires et suffisants) a été isolé grâce à l'utilisation de calcul intensif via le logiciel d'exploration de modèles OpenMole [SCHMITT et al., 2014], ce qui était un résultat impossible à atteindre de manière analytique pour un tel type de modèle complexe. Les progrès techniques d'OpenMole [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013] ont été menés simultanément avec les avances théoriques et empiriques. Les avancées épistémologiques ont également été cruciales dans ce cadre, comme REY le développe dans [REY-COYREHOURCQ, 2015], et de nouveaux concepts comme la modélisation incrémentale [COTTINEAU, CHAPRON et REUILLO, 2015] ont été découverts, avec de puissantes applications concrètes : [COTTINEAU, 2014] l'applique sur le système

de villes soviétique et isole les processus socio-économiques dominants, par un test systématique des hypothèses thématiques et des fonctions d'implémentation. Des directions pour le développement de telles pratiques de Modélisation et Simulation en géographie quantitative ont récemment été introduits par BANOS dans [BANOS, 2013]. Il conclut par neuf principes⁴, parmi lesquels on peut citer l'importance de l'exploration intensive des modèles computationnels et l'importance du couplage de modèles hétérogènes, qui sont avec d'autre principes tel la reproductibilité au centre de l'étude des systèmes complexes géographiques selon le point de vue décrit précédemment. Nous nous positionnons dans l'héritage de cette ligne de recherche, travaillant de manière conjointe sur les aspects théoriques, empiriques, épistémologiques et de modélisation.

C : (Florent) point intéressant, mais avant de prendre position pour intégration théorique/empirique, il faut qu'on comprenne pourquoi compliqué à faire (même si hyper riche, déjà des éléments en l'état dans le manuscrit

QUESTION DE RECHERCHE

C : (Florent) logique de dire cela à ce stade mais pas dans manuscrit final La question de recherche et les objets précis sont délibérément flous pour l'instant, puisque nous postulons que la construction d'une problématique ne peut être dissociée de la production d'une théorie correspondante. De manière réciproque, il n'y a aucun sens à poser des questions sorties de nulle part, sur des objets qui ont été seulement partiellement ou brièvement définis. Notre question préliminaire pour entrer dans le sujet, qu'on peut obtenir à partir de cas concrets comme l'anecdote introductory ou la revue de littérature préliminaire, est la suivante :

Comment définir les systèmes territoriaux, et les échelles et ontologies associées, dans une théorie cohérente, innovante et informative sur les processus sous-jacents ? **C : (Florent) très général et fausse question !**

C : (Arnaud) Très général, à voir si se tient

Il s'agit bien sûr d'une fausse question à ce stade, mais qui est toujours utile pour diriger la compréhension globale et le lecteur soucieux d'une démarche linéaire classique.

En effet, une caractéristique fondamentale des systèmes territoriaux est leur nature spatio-temporelle, qui est contenue dans leur dynamiques spatio-temporelles. La notion de *processus* au sens de [*Hypergeo*] capture de plus les relations causales entre composantes de ces dynamiques, et est ainsi une approche intéressante pour une compréhension voire explication de ces systèmes. L'échelle doit être comprise

⁴ Je me rappelle RENÉ DOURSAT insister pour la recherche du dernier commandement de BANOS

ici au sens opérationnel (caractéristiques physiques) end l'*ontologie* comme les objets réels étudiés⁵. Notre question peut être vue grossièrement comme la recherche de théories et modèles qui révèlent des processus impliqués dans des systèmes complexes contenant aux moins des établissements humains, ce dernier point étant crucial pour la construction d'une problématique convergente plutôt que de se perdre dans des propositions irréalistes et non constructives qui pourrait aller de comprendre tout du cerveau (qui peut être vu comme une brique élémentaire des systèmes territoriaux qui émergent des interactions sociales) à l'écosphère qui inclut aussi les systèmes territoriaux. Ces systèmes spatiaux, que nous préciserons comme *systèmes territoriaux*

C : (Florent) ok bien de préciser cela, mais peut être plus spécifique que de rappeler dimension territoriale (par ex. introduire bifurcations)

CONTENU

This provisory Memoire is organized the following way. A first part with four chapters sets the thematic, theoretical and methodological background. The study of geographical systems implies, because of their complexity, a subtle combination of Theoretical constructions and Empirical Analysis, either in an inductive reasoning or in a didactic constitution of knowledge. The first part aims to approach our subject from the theoretical and methodological point of view, and rather as a *necessary foundation* shall be understood as a body of knowledge *coevolving* with Empirical and Modeling Parts. A linear reading is not necessarily the best way to deeply perceive the implications of theory on empirical and modeling experiments and reciprocally. Some methodological developments are necessary but explicit reference will be done when it will be the case. A first chapter starts from the provisory research question given above and frames from a thematic point of view geographical objects and processes to be studied, resulting in precise research questions. The scene is set up for the construction of our theoretical background in a second chapter, that consists in a geographical theory for territorial systems on the one hand and in an epistemological theory of socio-technical systems **C : (Florent) c'est quoi ?** modeling that frames our approach at a meta-level. **C : (Florent) sens ?** We then develop methodological considerations on diverse questions implied by theory and required for modeling.

⁵ cet usage de la notion d'*ontologie* biaise naturellement la recherche vers des paradigmes de modélisation puisque qu'elle est proche de celle utilisée dans [LIVET et al., 2010], mais nous prenons la position (développée en détails plus loin) de comprendre toute construction scientifique comme un *modèle*, rendant la frontière entre théories et modèles moins pertinentes que pour des visions plus classiques. Toute théorie doit faire des choix sur les objets décrits, leur relations et les processus impliqués, et contient donc une *ontologie* dans ce sens.

Finally, a chapter of quantitative epistemology finishes to pave the way for modeling directions, unveiling literature gaps precisely linked to our question. A second part develops results obtained from empirical analysis and modeling experiments, along with on-going and planned projects in these fields. It first present empirical analysis aimed at identifying stylized facts. Toy-models of urban growth are then proposed, followed by an example and propositions for more complex models. The third part constructs our research objective for the remaining part of our project and sets a corresponding roadmap. Appendices contain non-digest important parts of our work such as models implementation architecture and details and specific tools developed for a reproducible research workflow.

SUR LA LECTURE LINÉAIRE

TODO : *expliquer notre position sur la difficulté d'une présentation linéaire, au delà de faire la synthèse. // bon bouquins y arrivent ? y réfléchir*

Première partie

FOUNDATIONS

This part set up foundations, constructing our research precise subject and questions from a thematic point of view, completed with a theoretical construction for framing at thematic and epistemological levels. We also provide methodological digressions, and a quantitative epistemological analysis completing the manual state of the art. **C : (Arnaud) ça s'appelle lire**

INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

Si la question de la priorité de l'œuf sur la poule ou de la poule sur l'œuf vous embarrasse, c'est que vous supposez que les animaux ont été originaiement ce qu'ils sont à présent.

- DENIS DIDEROT [DIDEROT, 1965]

Cette analogie est idéale pour introduire les notions de causalité et de processus dans les systèmes territoriaux. En voulant traiter naïvement des questions similaires à notre question de recherche préliminaire, certains ont qualifiés les causalités au sein de systèmes complexes comme un problème “de poule et œuf” **C : (Florent) parler à ce stade de la controverse Offner 93 C : (Arnaud) :** si un effet semble causer l'autre et réciproquement, comment est-il possible d'isoler les processus correspondants ? Cette vision est souvent présente dans les approches réductionnistes qui ne postulent pas une complexité intrinsèque au sein des systèmes étudiés. L'idée suggérée par DIDEROT est celle de *co-evolution* qui est un phénomène central dans les dynamiques évolutionnaires des Systèmes Complexes Adaptatifs comme HOLLAND élabore dans [HOLLAND, 2012]. Il fait le lien entre la notion d'émergence (ignorée dans les approches réductionnistes) **C : (Florent) la encore très epistemo, renforcer connaissance empirique de ces interactions particulières et en faire état ici , en particulier l'émergence de structures à une plus grande échelle par les interactions entre agents à une échelle donnée, en général concrétisée par un système de limites, qui devient cruciale pour la co-évolution des agents à toutes les échelles : l'émergence d'une structure sera simultanée avec une autre, chacune exploitant leur interrelations et environnements générés conditionnés par le système de limites. Nous explorerons ces idées pour le cas des systèmes territoriaux par la suite.** **C : (Florent)c'est seulement là que tu dis que les syst. territoriaux sont une déclinaison des questionnement précédents**

Ce chapitre introductif est destiné à poser le cadre thématique, le contexte géographique sur lesquels les développements suivants se baseront. Il n'est pas supposé être compris comme une revue de littérature exhaustive ni comme les fondations théoriques fondamentales de notre travail (le premier point étant l'objet du chapitre ?? tandis que le second sera traité plus tôt dans le chapitre 9), mais plutôt

comme une construction narrative ayant pour but d'introduire nos objets et positions d'étude, **C : (Florent) le faire plutot que l'annoncer** afin de construire naturellement des questions de recherche précises.

★ ★

★

Ce chapitre est entièrement inédit.

1.1 RÉSEAUX ET TERRITOIRES

1.1.1 Une circularité naturelle

TERRITORIALITÉ HUMAINE Une entrée possible dans l'ensemble des objets géographiques que nous proposons d'étudier est la notion de territoire. **C : (Florent) et un objet de recherche en lui même** En Ecologie, un territoire correspond à l'étendue spatiale occupée par un groupe d'agent ou plus généralement un écosystème. Les *Territoires Humains* sont extrêmement plus complexes de par l'importance de leur représentations sémiotiques, qui jouent un rôle significatifs dans l'émergence des constructions sociétales. **C : (Florent)pas besoin ni interet de se positionner sur emergence des societes** Selon RAFFESTIN dans [RAFFESTIN, 1988], la *Territorialité Humaine* est "la conjonction d'un processus territorial avec un processus informationnel", ce qui implique que l'occupation physique et l'exploitation de l'espace par les sociétés humaines n'est pas dissociable **C : (Florent) ou est complémentaire?** des représentations (cognitives et matérielles) de ces processus territoriaux, qui influent en retour leur évolution. En d'autres termes, à partir de l'instant où les constructions sociales déterminent la constitution des établissements humains, les structures sociales abstraites et concrètes joueront un rôle dans l'évolution des systèmes territoriaux, par exemple à travers la propagation d'informations et de représentations, par des processus politiques, ou encore par la correspondance effective entre territoire vécu et territoire perçu. **C : (Florent)donner exemples concrets serait pédagogique (ex metropole grd paris, cf articles)** Bien que cette approche ne donne pas de conditions explicites pour l'émergence d'un système séminal d'établissements agrégés (c'est à dire l'émergence des villes), **C : (Florent)pourquoi cette interrogation particulière?** elle insiste sur leur rôle comme lieu de pouvoir et de création de richesse au travers des échanges. Mais la ville n'a pas d'existence sans son hinterland et le système territorial peut difficilement être résumé par ses villes, comme un système de villes. En se restreignant à ce sous-système, il y a toutefois compatibilité entre la théorie de territoires de RAFFESTIN et la théorie évolutive des villes de PUMAIN [PUMAIN, 2010], qui interprète les villes comme des systèmes complexes dynamiques auto-organisés, **C : (Arnaud) self-organized?** qui agissent comme des médiateurs du changement social : par exemple, les cycles d'innovation s'initialisent au sein des villes et se propagent entre elles. **C : (Florent)tres pertinent bien sur, mais va aborder la question de l'innovation dans la thèse?** Les villes sont ainsi des agents **C : (Arnaud) entities?** compétitifs qui co-évoluent (au sens donné précédemment). Le système territorial peut ainsi être compris comme une structure sociale organisée dans l'espace, qui comprend ses artefacts concrets et abstraits. Une étendue spatiale imaginaire

avec des ressources potentielles qui n'aurait jamais connu de contact avec l'humain ne pourra pas être un territoire si elle n'est pas habitée, imaginée, vécue, exploitée, même si ces ressources pourraient être potentiellement exploitée le cas échéant. En effet, ce qui est considéré comme une ressource (naturelle ou artificielle) dépendra de la société (par exemple de ses pratiques et de ses capacités technologiques). [DI MEO, 1998] procède à une analyse historique des différentes conceptions de l'espace (qui aboutissent entre autre à l'espace vécu, l'espace social et l'espace classique de la géographie) et montre comment leur combinaison forme ce que RAFFESTIN décrit comme territoires. Un aspect central des établissements humains qui a une longue tradition d'étude en géographie, et qui est directement relié à la notion de territoire, est celui des *réseaux*. Nous allons voir comment le passage de l'un à l'autre est inévitable et leur définition indissociable.

C : (Florent)structure générale de l'argumentaire tb, mais devrait expliquer plus en détail ce qu'on appelle réseau (avant de détailler les différents réseaux réel/virtuel, les réseaux ont une inscription spatiale

UNE THÉORIE TERRITORIALE DES RÉSEAUX Nous paraphrasons DUPUY dans [DUPUY, 1987] lorsqu'il propose des éléments pour une "théorie territoriale des réseaux" basée sur le cas concret d'un réseau de transport urbain. Cette théorie présente les *réseaux réels* (i.e. les réseaux concrets, incluant les réseaux de transport) comme la matérialisation de *réseaux virtuels*.

C : (Florent) dans un second temps seulement, à ce stade "de qui viennent les réseaux" n'est pas une question cruciale, c'est la question réseau/espace/human settlements qui doit être au cœur

Plus précisément, un territoire est caractérisé par de fortes discontinuités spatio-temporelles induites par la distribution non-uniforme des agents

C : (Arnaud) Ontology et des ressources. Ces discontinuités induisent naturellement un réseau de "projets transactionnels"

C : (Florent) pourquoi guillemets ? qui peuvent être compris comme des interactions potentielles entre les éléments du système territorial (agents et/ou ressources). Par exemple, de nos jours les actifs se doivent d'accéder à la ressource qu'est l'emploi, et des échanges économiques s'effectuent entre les différents territoires spécialisés dans les productions de différents types. En tout temps des interactions potentielles ont existé¹ Le réseau d'interaction potentiel est concrétisé quand l'offre s'adapte à la demande, et résulte en la combinaison de contraintes économiques et géographiques avec les motifs de demande, de manière non-linéaire via des agents qu'on peut désigner comme *opérateurs*. Un tel processus est loin d'être immédiat, et conduit à de

¹ même quand le nomadisme devait encore être la règle, des réseaux d'interactions potentielles dynamiques dans l'espace ont pu exister, mais devaient avoir moins de chance de se matérialiser en des routes matérielles.

forts effets de non-stationarité et de dépendance au chemin C : (Florent)Une strategie à adopter serait d'abord de decrire de facon basique, avec exemples concrets, la complexité des interactoins réseaux/espace/settlements, puis de rappeler CS et proprietes, puis de decrire lesquelles de ces propriétés presentes dans ces interactions, lequelles modèles vont essayer de reproduire et pquois. : l'extension d'un réseau existant dépendra de la configuration précédente, et selon les échelles de temps impliquées, la logique et même la nature des opérateurs peut avoir évolué. RAFFESTIN souligne dans sa préface de [OFFNER et PUMAIN, 1996] qu'une théorie géographique articulant espaces, réseaux et territoires n'a jamais été formulée de manière cohérente. C : (Florent)redire les ecueils qui sont perçus par Raffestin Il semble que c'est toujours le cas aujourd'hui, même si la théorie évoquée ci-dessus semble être un bon candidat bien qu'elle reste à un niveau conceptuel. La présence d'un territoire humain implique nécessairement la présence de réseaux d'interactions abstraites et de réseaux concrets utilisés pour transporter les individus et les ressources (incluant les réseaux de communication puisque l'information est une ressource essentielle). Selon le régime dans lequel le système considéré se trouve, le rôle respectif du réseau peut être radicalement différent. Selon DURANTON [DURANTON, 1999], les villes pré-industrielles étaient limitées en croissance de par les limitations des réseaux de transport. Les progrès technologiques ont permis de les surmonter C : (Florent)trop simplificateur et à mené à la prépondérance du marché foncier dans la formation des villes (et par conséquent un rôle des réseaux de transport qui déterminent les prix par l'accessibilité), et plus récemment à une importance croissante des réseaux de télécommunication ce qui a induit une "tyrannie de la proximité" puisque la présence physique n'est pas remplacable par une communication virtuelle. Cette approche territoriale des réseaux semble naturelle en géographie, puisque les réseaux sont étudiés conjointement avec des objets géographiques auxquels est associée une théorie, en opposition à la science des réseaux qui étudie brutalement les réseaux spatiaux avec peu de fond thématique [DUCRUET et BEAUGUITTE, 2014]. C : (Florent)derniere phrase pas claire C : (Arnaud) Ajouter noms ? (biblio ?

DES RÉSEAUX QUI FAÇONNENT LES TERRITOIRES ? Cependant les réseaux ne sont pas seulement une manifestation matérielle de processus territoriaux, mais jouent également leur rôle dans ces processus comme leur évolution peut influencer l'évolution des territoires en retour. Dans le cas des *réseaux techniques*, une autre désignation des réseaux réels donnée dans [OFFNER et PUMAIN, 1996], de nombreux exemples de tels retroactions peuvent être mis en évidence : l'interconnexion des réseaux de transport permet des motifs de mobilité multi-échelles, C : (Florent)chose plus basiques à dire

en premier (favorise croissance urbaine) formant ainsi le territoire vécu. A une plus petite échelle, des changements de l'accessibilité peuvent induire l'adaptation d'un espace fonctionnel urbain. Il émerge alors une difficulté intrinsèque : **C : (Florent) TB mais en parler avant, c'est cela le cœur** il est loin d'évident d'attribuer des mutations territoriales à une évolution du réseau and réciproquement la matérialisation d'un réseau à des dynamiques territoriales précises. Revenir à la citation de Diderot devrait aider à ce point, au sens où il ne faut pas considérer le réseau ni les territoires comme des systèmes indépendants qui s'influencerait mutuellement par des relations causales, mais comme des composantes fortement couplées d'un système plus large. La confusion autour de possibles relations causales simples a nourri un débat scientifique encore actif aujourd'hui. Les méthodologies pour identifier ce qui est nommé *effets structurants* des réseaux de transport ont été proposées par les planificateurs dans les années 1970 [BONNAFOUS et PLASSARD, 1974; BONNAFOUS, PLASSARD et SOUM, 1974]. **C : (Florent)TB; c'est toujours un débat d'actualité (ok dit)** Il aura fallu un certain temps pour un positionnement critique sur l'usage non raisonné et decontextualisé de ces méthodes par les planificateurs et les politiques généralement pour justifier technocratiquement des projets de transports. Cela a été fait en premier par OFFNER dans [OFFNER, 1993]. Récemment un édition spéciale du même journal sur ce débat [L'ESPACE GÉOGRAPHIQUE, 2014] a rappelé d'une part que les mauvaises interprétations et les mauvais usages étaient encore largement présent aujourd'hui dans les milieux opérationnels de la planification comme [CROZET et DUMONT, 2011a] confirme, et d'autre part qu'il faudrait encore une certaine quantité de progrès scientifique pour comprendre en profondeur les relations entre réseaux et territoires. PUMAIN souligne que des travaux récents ont révélé des effets systématiques sur de très longues échelles temporelles (comme e.g. le travail de BRETAGNOLLE sur l'évolution des chemins de fer, qui montre une sorte d'effet structurel sur la nécessité de connexion au réseau des villes, afin de rester actives, mais qui n'est ni suffisant ni totalement causal). **C : (Florent)développer ce genre de catégories macro c'est très intéressant** A un niveau macroscopique des motifs typiques d'interaction émergent, mais les trajectoires microscopiques du systèmes sont essentiellement chaotiques : la compréhension des dynamiques couplées dépend fortement de l'échelle considérée. A une petite échelle il est peu raisonnable de vouloir montrer des comportement systématiques, comme le rappelle OFFNER. Par exemple, sur des territoires de montagne français comparables, [BERNE, 2008] montre que les réactions à un même contexte d'évolution du réseau de transport peut mener à des réactions territoriales très diverses, certains trouvant de forts bénéfices par la nouvelle connectivité, d'autres au contraire devenant plus fermés. Ces retroactions potentielles des réseaux sur les territoires n'agit

pas nécessairement sur des composantes concrètes : CLAVAL montre dans [CLAVAL, 1987] que les réseaux de transport et de communication contribuent à la représentation collective d'un territoire en agissant sur un sentiment d'appartenance. **C : (Florent) la encore de second ordre, a ressortir pour lutetia**

SYSTÈMES TERRITORIAUX Ce voyage des territoires aux réseaux, et retour, nous permet d'esquisser une définition préliminaire d'un système territorial sur laquelle se basera les considérations théoriques suivantes. **C : (Florent)si c'est autant au coeur, présenter avant Comme nous avons mis en exergue le rôle des réseaux, la définition se doit de les prendre en compte.**

Définition provisoire. *Un Système Territorial est un territoire humain auquel peuvent être associés à la fois un réseau d'interactions et un réseau réel. Les réseaux réels sont une composante à part entière du système, jouant dans les processus d'évolution, au travers de multiples retroactions avec les autres composantes à plusieurs échelles spatiales et temporelles.*

C : (Florent) feedback : propriété, pas def ; plus une axiomatique qu'une demo ?

Cette lecture des systèmes territoriaux est conditionnée à l'existence des réseaux et pourrait écarter certains territoires humains, mais il s'agit d'un choix délibéré justifié par les considérations précédentes, et qui précise notre sujet vers l'étude des interactions entre réseaux et territoires. **C : (Florent) formulé comme ça, on peut penser que network pas inclus dans territoire**

1.1.2 Réseaux de Transport

LA PARTICULARITÉ DES RÉSEAUX DE TRANSPORT Déjà évoqués dans le cas des effets structurants des réseaux, les réseaux de transports jouent un rôle déterminant dans l'évolution des territoires. **C : (Florent) donc souscrit à théorie des effets structuraux causaux ?** Même si d'autres types de réseaux sont également fortement impliqués dans l'évolution des systèmes territoriaux (voir e.g. les débats sur l'impact des réseaux de communication sur la localisation des activités économiques), les réseaux de transport conditionnent d'autres types de réseaux (logistique, échanges commerciaux, interactions sociales concrètes pour donner quelques exemples) and semblent dominer dans les motifs d'évolution territoriale, en particulier dans nos sociétés contemporaines qui sont devenues dépendantes des réseaux de transport [BAVOUX et al., 2005]. Le développement du réseau français à grande vitesse est une illustration pertinente de l'impact des réseaux de transport sur les politiques de développement territorial. Présenté comme une nouvelle ère de transport sur rail, une planification par le haut de lignes totalement nouvelles **C : (Florent)**

et x2 speed a été présenté comme central pour le développement [ZEMBRI, 1997]. Le manque d'intégration de ces nouveaux réseaux avec l'existant et avec les territoires locaux est à présent observé comme une faiblesse structurelle et des impacts négatifs sur certains territoires ont été prouvés [ZEMBRI, 2008]. Une revue faite dans [BAZIN et al., 2011] confirme qu'aucune conclusion générale sur des effets locaux d'une connection à une ligne à grande vitesse ne peut être tirée, **C : (Florent) va trop vite beaucoup de cas différents (ouest/rhone/strasbourg/lille/rhin-rhone)** bien que ce sésame garde une place conséquente dans les imaginaires des élus. Ces exemples illustrent comment les réseaux de transport peuvent avoir des effets à la fois directs et indirects sur les dynamiques territoriales. La planification intégrée, au sens d'une planification coordonnée entre les infrastructures de transport et le développement urbain, considère le réseau comme une composante déterminante du système territorial. Les Villes Nouvelles parisiennes sont un tel cas qui témoigne de la complexité de ces actions de planification qui le plus souvent ne mène pas au effets initialement désirés [OSTROWETSKY, 2004]. Des projets récents comme [L'HOSTIS, SOULAS et WULFHORST, 2012] ont tenté d'implémenter des idées similaires, mais il manque pour l'instant de recul pour juger de leur succès à produire un territoire effectivement intégré. **C : (Florent) dans le detail, quels sont les ordres de grandeur des temps pour que les réseaux puissent avoir un effet?** Les réseaux de transports sont dans tous les cas au centre de ces approches des territoires urbains. Nous nous concentrerons par la suite sur les réseaux de transport **C : (Florent)tous?** pour toutes ces raisons évoquées ici.

DÉCONSTRUIRE L'ACCESSIBILITÉ [MILLER, 1999] on three different way to approach accessibility : time-geography and constraints, user utility based measures, and transportation time. It derives measures for each in perspective of WEIBULL's axiomatic frameworks and reconcile the three in a way.

La notion d'accessibilité surgit rapidement lorsqu'on s'intéresse aux réseaux de transport. Basée sur la possibilité d'accéder un lieu par un réseau de transport (pouvant prendre en compte la vitesse, la difficulté de se déplacer), elle est généralement définie comme un potentiel d'interaction spatiale² [BAVOUX et al., 2005]. Cet objet est souvent utilisé comme un outil de planification ou comme une variable explicative de localisation des agents par exemple. **C : (Florent) dire d'abrd à quoi peut servir** Il faut cependant rester prudent sur son usage inconditionnel. Plus précisément, il peut s'agir d'une construction qui ignore une partie conséquente des dynamiques ter-

² et souvent généralisée comme une *accessibilité fonctionnelle*, par exemple les emplois accessibles aux actifs d'un lieu. Les potentiels d'interaction spatiaux s'exprimant dans les lois de gravité peuvent aussi être compris de cette façon.

ritoriales. La mystification C : (Florent) trop fort, Hadri montre que étude et prod de l'infra sont pas indep, mais pas de myst C : (Arnaud) Contexte français de la notion de *mobilité* a été montrée par COMMENGES dans [COMMENGES, 2013b], qui révèle que la majorité des débats sur la modélisation de la mobilité et les notions correspondantes était majoritairement construites de manière ad-hoc par les administrateurs de transports issus du *Corps des Ponts* C : (Florent) lecture trop rapide qui importaient brutalement les outils et méthodes des Etats-Unis sans adaptation ni reflexion adaptée au contexte français. L'accessibilité pourrait de même être une construction sociale et n'avoir que peu de fondement théorique, puisqu'il s'agit en grande partie d'un outil de modélisation et de planning. Les débats récents sur la planification du *Grand Paris Express* [MARGIN, 2014], C : (Florent) intéressant : à creuser cette nouvelle infrastructure de transport métropolitaine planifiée pour les vingts prochaines années, a révélé l'opposition entre une vision de l'accessibilité comme un droit pour les territoires désavantagés, contre l'accessibilité comme un moteur du développement économique pour des zones déjà dynamiques, les deux étant difficilement compatibles car correspondent à des couloirs de transport très différents. De tels problèmes opérationnels confirment la complexité du rôle des réseaux de transports dans les dynamiques des systèmes territoriaux, et nous devrons donner dans notre travail des éléments de réponse pour une définition de l'accessibilité qui intégrerait les dynamiques territoriales intrinsèques.

ECHELLES ET HIERARCHIES Un aspect incontournable des réseaux de transport que nous devrons prendre en compte dans nos développements futurs est la hiérarchie. Les réseaux de transport sont par essence hiérarchique, dépendant des échelles dans lesquelles ils sont intégrés. [LOUF, ROTH et BARTHELEMY, 2014] montre empiriquement des propriétés de loi d'échelle pour un nombre conséquent d'aires métropolitaines à travers la planète, et les lois d'échelle révèlent la présence de hiérarchie dans un système, comme pour la hiérarchie de taille dans les systèmes de villes exprimée par la loi de Zipf [NITSCH, 2005] ou d'autres lois d'échelle urbaines [ARCAUTE et al., 2013; BETTENCOURT et LOBO, 2015]. La topologie du réseau de transport a été montrée suivre de telles lois pour la distribution de ses mesures locales comme la centralité [SAMANIEGO et MOSES, 2008]. C : (Florent) tb mais comment relie à partie juste avant ? La hiérarchie semble jouer un rôle particulier dans les processus d'interaction, comme BRETAGNOLLE [BRETAGNOLLE, 2009a] a souligné une corrélation croissante dans le temps entre la hiérarchie urbaine et la hiérarchie du réseau pour le réseau ferroviaire français, C : (Florent) tb mais séparé entre réseau et pop ; pourquoi pas regarder la hiérarchie de l'access ? marqueur de retroactions positives entre le rang urbain et la centralité

de réseau. Différents régimes dans le temps et l'espace ont été identifiés : pour l'évolution du réseau ferroviaire français e.g., une première phase d'adaptation du réseau à la configuration urbaine existante a été suivie par une phase de coévolution i.e. au sens où les relations causales sont devenues difficiles à identifier. L'impact de la contraction de l'espace-temps par les réseaux sur le potentiel de croissance des villes avait déjà été montré pour l'Europe par des analyses exploratoires dans [BRETAGNOLLE, PUMAIN et ROZENBLAT, 1998]. L'évolution du réseau ferroviaire aux Etats-unis a suivi une dynamique bien différente, sans diffusion hiérarchique, donnant forme localement à la croissance urbaine. **C : (Florent) un peu rapide mais dans l'autre sens : cela n'a pas marché partout mais contexte particulier de la conquête de l'ouest est intéressant à souligner** Cela met l'emphase sur la présence de dépendance au chemin **C : (Florent) en parler avant si c'est le cœur du projet** pour les trajectoires des systèmes urbains : la présence en France d'un système préalable de villes et de réseau (routes postales) a fortement influencé le développement du réseau ferré, tandis que son absence aux Etats-unis a conduit à une histoire complètement différente. Une question ouverte est si des processus génériques sont implicites aux deux évolutions, chacun correspondant à des réalisations différentes avec des conditions initiales et des méta-paramètres différentes (des *régimes* différents au sens des transitions des systèmes de peuplement introduites dans le projet de recherche courant ANR TransMonDyn, puisque une transition peut être comprise comme un changement de stationnarité des méta-paramètres **C : (Florent) trop rapide ce n'est pas compréhensible en l'état** d'une dynamique générale). En termes de systèmes dynamiques, cela revient à se demander si les dynamiques de attracteurs **C : (Florent) en considérant qu'ils existent** (composantes à grande échelle temporelle) obéissent à des équations similaires que la position et nature des attracteurs pour un système dynamique stochastique qui donnent son régime courant, en particulier si le système est dans un état local divergent (exposant de Liapounov local positif) ou en train de converger vers des mécanismes stables [SANDERS, 1992]. Pour répondre à cette question en même temps que l'isolation des processus de coévolution pour ce régime, [BRETAGNOLLE, 2009a] propose la modélisation comme élément de réponse constructif. Nous verrons dans le chapitre suivante comme la modélisation peut être source de connaissance à propos de processus territoriaux.

sur la mobilité : nos questionnements à une autre échelle ? cf [FUSCO, 2004] relations causales

1.1.3 *Interactions entre Réseaux et Territoires*

At this state of progress, we have naturally identified a research subject that seems to take a significant place in the complexity of territo-

rial systems, that is the study of interactions between transportation networks and territories. In the frame of our preliminary definition of a territorial system, this question can be reformulated as the study of networked territorial systems with an emphasize on the role of transportation networks in system evolution processes.

C : (Florent) ok : à quelles échelles de temps et d'espace se place t'on (même un intervalle)

- ici donner des exemples concrets -

Gaelle Lesteven Metro toulouse

C : (Florent) aéroport MCR : Ciudad Real

1.2 DE PARIS À ZHUHAI

1.2.1 *Le Grand Paris : histoire et enjeux*

La région parisienne est une bonne illustration de la complexité des interactions entre réseaux de transports et territoires, au cours du temps et à l'échelle intermédiaire d'une région métropolitaine globalement mono-centrique.

[GILLI et OFFNER, 2009] propose en 2009 un diagnostic de la situation institutionnelle de la région parisienne, et des pistes pour une approche couplée entre gouvernance et aménagement. La préfiguration de "l'instauration d'un acteur collectif métropolitain" correspond à la métropole du Grand Paris qui sera inaugurée 7 ans plus tard

1.2.2 *Le Delta de la Rivière des Perles : nouveaux régimes urbains et Mega-City Regions*

TODO : some “comparable” maps would be useful : ask Chenyi most precise data on PRD : territorial variables and transportation networks ?

Parler du pont et des bifurcations induites (cf intro chap 5)

Si la notion de megalopolis peut être tracée jusqu'à GOTTMANN [GOTTMANN, 1964], et qu'elle est à l'origine de celle de Mega-city Region consacrée par HALL [HALL et PAIN, 2006], il est clair que cette dernière est toujours plus d'actualité avec l'apparition récente de nouveaux régimes, notamment par l'urbanisation croissante dans des pays à forte croissance et en mutation très rapide comme la Chine [SWERTS et DENIS, 2015].

1.2.3 *Comparabilité des études de cas*

1.3 ELEMENTS DE TERRAIN

1.3.1 Une Experience en Observation Flottante

Si le diable est dans les détails, les systèmes de transport entre autres sont l'allégorie de cette adage. Ce que certains appellent détail contient la majorité de l'information pour d'autres. Logiquement enfermés dans une bulle scientifique, malgré toutes les volontés développées en introduction, on tâchera de rester conscient de la nature et la portée de la connaissance produite ici. Ce que nous pourrions appeler détail, lors de l'étude de l'accessibilité d'un réseau de transport par exemple, tel des impressions ressenties par les usagers ou les relations sociales induites par les situations découlant des dynamiques du systèmes, seront le centre du questionnement pour un anthropologue ou sociologue. Une telle connaissance, qui trouverait certainement une place dans nos problématiques, est hors de notre portée de par l'absence de *terrain* de longue durée. Nous proposons toutefois ici d'ébaucher une entrée qualitative d'un certain type, pour suggérer une façon de compléter nos connaissances.

L'entrée prise suit la méthode *d'observation flottante*, introduite à l'interface de l'anthropologie et la sociologie par [PÉTONNET, 1982], avec l'ambition de fonder une anthropologie urbaine. Il ne s'agit pas exactement de la même idée que l'anthropologie de l'espace de Choay Répondant à un besoin de mouvement que le sédentaire éprouve facilement, le chercheur se place au centre du processus de production de connaissances, nous citons, en "rest[ant] en toute circonstance vacant et disponible, à ne pas mobiliser l'attention sur un objet précis, mais à la laisser flotter afin que les informations la pénètrent sans filtre, sans a priori, jusqu'à ce que des points de repère, des convergences, apparaissent et que l'on parvienne alors à découvrir des règles sous-jacentes". Sans s'y méprendre et considérer la méthode comme une négligence méthodologique, nous y voyons une opportunité d'un accès rapide et à faible coût dans le monde du qualitatif, tout en restant conscient de sa portée très limitée. La disposition d'esprit peut être rapprochée de la philosophie La méthode peut servir d'étude préliminaire pour fixer des protocoles et grilles précises d'entretien : elle est par exemple utilisée justement au sujet du transport par [ALBA et AGUILAR, 2012].

Les mouvements pendulaires à échelle moyenne sont nécessairement vécus d'une façon particulière en comparaison à d'autres lieux géographiques et à d'autres échelles sur le même lieu. Et si une façon d'appréhender des faits stylisés particuliers était alors d'effectuer l'analogie d'une étude de perturbation sur le système, mais en prenant comme référentiel l'observateur lui-même ? Il s'agirait de faire porter un choc sur une situation "d'équilibre", puis de se laisser flotter au gré du courant pour appréhender la réaction et certains mé-

canismes qu'il aurait été difficile de considérer en suivant sa routine. Une expérience naturelle causée par une perturbation des transports (qui en région francilienne est bien courante) est un événement déclencheur de "naufrages" de l'observation, au sens où le chercheur peut capturer des situations et réactions individuelles particulières.

AU-DELÀ DU CHARLATANISME : SYSTÉMATISER LA MÉTHODE FLOTANTE Notre méthodologie est relativement simple : se laisser errer dans les transports en commun, avec ou sans but et de manière ou non aléatoire, mais en essayant sur chaque trajet de maximiser les opportunités de mise en situation ou de capture d'évènement. La répétition de l'expérience visera également à maximiser la couverture spatiale, temporelle, de situation. Une production traçable est nécessaire à chaque itération, qu'il s'agisse de description factuelle, de description perçue, de semi-synthèse

1.3.2 *Entretiens*

1.3.3 *Analyse Urbanistique*

Le ciel est gris et les visages fermés, Oxmo avait tristement raison, ce Soleil du Nord n'avait de lumière que le nom. L'initié ne saura s'y tromper et ressentira au fond de lui-même cette banale routine d'un aller-retour quotidien en RER. Il ne cherchera ni à maudire les planifications successives dont les stratifications temporelles ont laissé décanter cette organisation territoriale incongrue, ni à se prendre à rêver d'une trajectoire de vie alternative puisque choisir c'est un peu mourir et qu'il ne se sent pas une âme de Phoenix aujourd'hui. Peut être que la beauté de la ville est finalement dans ces tensions qui la façonnent à tous les niveaux et dans tous les domaines, ces paradoxes qui deviennent cadre de vie au point d'asséner quotidiennement une vérité. Cette philosophie de couloir de métro, le francilien en fait son cheval de bataille car après tout s'il vit en ville il doit bien la connaître. Encore un rail cassé sur le A, "tout cela est mal géré, et ce réseau est mal conçu" vocifère un utilisateur journalier, s'improvisant expert en planification ; d'autres plus patients prennent leur mal en patience mais se présentent tout aussi connaisseurs d'une illusoire vision d'ensemble d'un territoire aux multiples visages. Ces usagers *sont* pourtant le système, de manière concrète à leur échelle d'espace et de temps, par induction et émergence aux échelles supérieures. La fourmi est supposée ne pas avoir conscience de l'intelligence collective dont elle est une des composantes fondamentales. Ils n'ont de la même manière que peu de perception de l'auto-désorganisation dont ils sont la source, peut-être la cause, et qui très sûrement subissent les désagréments de ses dynamiques. Se laisser flotter dans les transports franciliens est une expérience intemporelle. Presque thérapeutique parfois, quand l'un commence à perdre son optimisme quant à l'intérêt d'une vie urbaine, une excursion aléatoire en métro rappelle rapidement la richesse et la diversité qui sont un des plus grand succès des villes. C'est cette variété apparente de profils que le chercheur retiendra principalement de ces errements dont la méthodologie est de ne pas avoir de méthodologie, et il gardera à l'esprit qu'il n'existe pas d'échelle où un traitement spécifique de chaque objet géographiques n'est pas nécessaire : en quelque stations sur la ligne 4 le profil des quartiers et donc des usagers change profondément et souvent sans transition au moins trois fois, comme sur la ligne 13 nord où les motifs horaires soulignent d'autant plus de dures réalités socio-économiques qui sont en fait géographiques dans cet *espace produit* de la métropole. Lorsqu'il s'agit de modéliser, prendre en compte les limites de toute tentative de généralisation est d'autant plus cruciale comme chaque modèle est un équilibre fragile entre spécificité et générnicité.

ENCADRÉ : *Une expérience en observation flottante en région parisienne*

ENCADRÉ : *Une expérience en observation flottante, Guangdong, Zhubai*

CONCLUSION DU CHAPITRE

2

MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

TODO : Citer conversation avec JP Marchand (Theo Quant) : "Notre génération a compris qu'il y avait une co-évolution, la votre cherche à la comprendre". Retourner l'interviewer ?

★ ★

★

Ce chapitre est inédit pour sa première section ; reprend dans sa deuxième section

2.1 MODÉLISER LES INTERACTIONS

2.1.1 Modélisation en Géographie Quantitative

TODO : épistémologie des modèles équilibre/hors équilibre (pas faut dans positionnement épistémo, le faire ici)

La modélisation joue en Géographie Théorique et Quantitative (TQG) un rôle fondamental. CUYALA procède dans [CUYALA, 2014] à une analyse spatio-temporelle du mouvement de la Géographie Théorique et Quantitative en langue française et souligne l'émergence de la discipline comme une combinaison d'analyses quantitatives (e.g. analyse spatiale et pratiques de modélisation et de simulation) et de construction théoriques. **C : (Florent) cela remonte à quand ? appliqué à quels champs ?** L'intégration de ces deux composantes permet la construction de théories à partir de faits stylisés empiriques, qui produisent à leur tour des hypothèses théoriques pouvant être testées sur les données empiriques. Cette approche est née sous l'influence de la *New Geography* dans les pays Anglo-saxons et en Suède. Une histoire étendue de la genèse des modèles de simulation en géographie est faite par REY dans [REY-COYREHOURCQ, 2015] avec une attention particulière pour la notion de validation de modèles. L'utilisation de ressources de calcul pour la simulation de modèles est antérieur à l'introduction des paradigmes de la complexité, remontant à HÄGERTRAND **C : (Florent) AB, conceptuel, pas computationnel** **C : (Arnaud) Hagerstrand NON** et FORRESTER, **C : (Arnaud) Forrester ≠ géographe** pionniers des modèles d'économie spatiale inspirés par la cybernétique. Avec l'augmentation des potentialités de calcul, des transformations épistémologiques ont également suivi, avec l'apparition de modèles explicatifs comme outils expérimentaux. REY compare le dynamisme des années soixante-dix quand les centres de calcul furent ouverts aux géographes à la démocratisation actuelle du Calcul Haute Performance (calcul sur grille à l'utilisation transparente, voir [SCHMITT et al., 2014] pour un exemple des possibilités offertes en terme de calibration et de validation de modèle, réduisant le temps de calcul nécessaire de 30 ans à une semaine - ces techniques jouent un rôle clé pour les résultats que nous obtiendrons par la suite), qui est également accompagnée par une évolution des pratiques [BANOS, 2013] et techniques [CHÉREL, COTTINEAU et REUILLON, 2015] de modélisation. La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [LIVET et al., 2010] rappelle la combinaison des domaines empirique, conceptuel (théorique) et de la modélisation, avec des rétroactions constructives entre chaque. Une modèle peut être un outil d'exploration pour tester des hypothèses, un outil empirique pour valider une théorie sur des jeux de données, un outil explicatif pour révéler des causalités et ainsi des processus internes au sys-

tème, un outil constructif pour construire itérativement une théorie conjointement avec celle des modèles associés. Ce sont des exemples de fonctions parmi d'autres : Varenne donne dans [VARENNE, 2010b] une classification raffinée des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein de systèmes complexes adaptatifs, et précisons encore notre question de recherche, qui s'intéressera aux modèles impliquant des interactions réseaux et territoires.

2.1.2 Modéliser les territoires et réseaux

Au sujet de notre question précise des interactions entre réseaux de transport et territoires, nous proposons un aperçu des différentes approches. Selon [BRETAGNOLLE, PAULUS et PUMAIN, 2002], "les idées des spécialistes de la planification cherchant à donner des définitions des systèmes de ville, depuis 1830, sont étroitement liées aux transformations des réseaux de communication". **C : (Florent) la question de la définition de la ville mérite une place plus grande** C'est en quelque sorte la prophétie auto-réalisatrice inversée, au sens où elle est déjà réalisée avant d'être formulée. Cela implique que les ontologies et les modèles correspondants proposés par les géographes et les planificateurs sont fortement liés aux préoccupations historiques courantes, ainsi forcément limités en portée et raisons. Dans une vision perspectiviste de la science [GIERE, 2010c] de telles limites sont l'essence de l'entreprise scientifique, et comme nous démontrerons en chapitre 9 leur combinaison et couplage dans le cas de modèles est une source de connaissance.

Modèles LUTI

Un partie importante de la littérature proposant des modélisations des interactions entre réseaux et territoires se trouve dans le domaine de la planification urbaine, avec les *modèles d'interaction entre usage du sol et transport (LUTI)*. Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines. Par exemple, du point de vue de l'Economie Urbaine, les propositions de modèle intégrés existent depuis un certain temps [PUTMAN, 1975]. La variété des modèles existants a conduit à des comparaisons opérationnelles [PAULLEY et WEBSTER, 1991; WEGENER, MACKETT et SIMMONDS, 1991]. Plus récemment, les avantages respectifs des approches statiques et dynamiques a été étudié par [KRYVOBOKOV et al., 2013]. **C : (Florent) ok mais spécifie des durées et échelles d'espace** Dans tous les cas, ce type de modèle opère généralement à des échelles temporelles et spatiales relativement faibles. [WEGENER et FÜRST, 2004] donne un état de l'art des études empiriques et de modélisation sur ce type d'approche des interactions entre usage du sol et transport. Le positionnement théorique est plutôt proche des disciplines de l'Economie, de la Pla-

nification et de la Sociologie, et relativement de nos raisonnements géographiques qui se veulent de comprendre également des processus sur le temps long. **C : (Florent) d'abord dresser le tableau des disciplines qui s'y intéressent, pourquoi et comment** Pas moins de dix-sept modèles sont comparés et classifiés, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement petites des simulations. Une revue complémentaire est faite par [CHANG, 2006], élargissant le contexte avec l'inclusion de classes plus générales de modèles, comme des modèles d'interactions spatiales (parmi lesquels l'attribution du traffic et les modèles à quatre temps), les modèles de planification basés sur la recherche opérationnelle (optimisation des localisations), les modèles microscopiques d'utilité aléatoire, et les modèles de marché foncier. Toutes ces techniques opèrent également à une petite échelle et considèrent au plus l'évolution de l'usage du sol. [IACONO, LEVINSON et EL-GENEIDY, 2008] couvre un horizon similaire avec une emphase supplémentaire sur les modèles à automates cellulaires d'évolution d'usage du sol et les modèles basés agent. Les modèles LUTI sont toujours largement étudiés et appliqués, comme par exemple [DE-LONS, COULOMBEL et LEURENT, 2008] qui est utilisé pour la région métropolitaine parisienne. La courte portée temporelle d'application de ces modèles et leur nature opérationnelle les rend utiles pour la planification, **C : (Florent) détailler ce que cela veut dire aidera certainement à mieux positionner par rapport au planning** ce qui est assez loin de notre souci d'obtenir des modèles explicatifs de processus géographiques.

Croissance du Réseau

La croissance de réseaux est pratiquée dans des entreprises de modélisation qui cherchent à expliquer de manière endogène **C : (Florent) de quel point de vue?** la croissance des réseaux de transport, généralement d'un point de vue *bottom-up*, i.e. en mettant en évidence des règles locales qui permettraient de reproduire la croissance du réseau sur de longues échelles de temps (souvent le réseau de rues). Les économistes ont proposés des modèles de ce type : [ZHANG et LEVINSON, 2007] passe en revue la littérature en économie de transports sur la croissance des réseaux dans le contexte d'une théorie endogène de la croissance [AGHION et al., 1998], rappelant les trois aspects principalement traités par les économistes sur le sujet, qui sont la tarification routière, l'investissement en infrastructures et le régime de propriété, et propose finalement un modèle analytique combinant les trois. [XIE et LEVINSON, 2009c] propose une revue étendue de la modélisation de croissance des réseaux, en prenant en compte d'autres champs : la géographie des transports a développé très tôt des modèles basés sur des faits empiriques mais qui se sont concentrés sur reproduire la topologie plutôt que sur les mécanismes selon [XIE et LEVINSON,

2009c]; les modèles statistiques sur des cas d'étude fournissent des conclusions très mitigées sur les relations causales entre offre et demande **C : (Florent) du coup ce n'est pas que pur réseau a priori A1 : todo : define what we mean by network**; les économistes ont étudié la production d'infrastructure à la fois d'un point de vue microscopique et macroscopique, généralement non spatiaux; la science des réseaux a produit des modèles jouet de croissance de réseau qui se basent sur des règles topologiques et structurelles plutôt que des règles se reposant sur des processus inspirés de faits réels. Une autre approche qui n'est pas mentionnée et que nous allons approfondir est la conception de réseau inspirée de la biologie. Nous donnons pour commencer des exemples d'études utilisant des concepts économiques ou géométriques pour modéliser la croissance de réseau. [YERRA et LEVINSON, 2005] montre avec un modèle économique basé sur des processus auto-renforçants et incluant une règle d'investissement basée sur l'attribution du trafic, que des règles locales sont suffisantes pour faire émerger une hiérarchie du réseau routier à usage du sol fixé. Une modèle très similaire donnée par [LOUF, JENSEN et BARTHELEMY, 2013] avec des fonctions coûts-bénéfices plus simples obtient une conclusion similaire. **C : (Florent) devrais rentrer plus dans le détail d'un ou deux modèles** Alors que ces modèles basés sur des processus cherchent à reproduire des motifs macroscopiques des réseaux (typiquement les lois d'échelle), les modèles d'optimisation géométrique cherchent à ressembler à des réseaux réels dans leur topologie. [BARTHÉLEMY et FLAMMINI, 2008] décrit un modèle basé sur une optimisation locale de l'énergie, mais ce modèle reste très abstrait et non validé. Le modèle de morphogenèse de [COURTAT, GLOAGUEN et DOUADY, 2011] qui utilise des potentiels locaux et des règles de connectivité, même s'il n'est pas calibré, semble reproduire de manière plus raisonnable des motifs réels des réseaux de rues. Un modèle très proche est décrit dans [RUI et al., 2013]. D'autres tentatives comme [DE LEON, FELSEN et WILENSKY, 2007; YAMINS, RASMUSSEN et FOGLER, 2003] sont plus proches de la modélisation procédurale [LECHNER et al., 2004; WATSON et al., 2008] et pour cette raison n'ont pas d'intérêt pour notre cas puisqu'ils peuvent difficilement être utilisés comme modèles explicatifs. **C (JR) : développer plus pourquoi la modélisation procédurale n'est pas satisfaisante : forme "fidèle" en général pas à la bonne échelle; penser qu'il s'agit de modèles de morphogenèse urbaine est une erreur grossière de POM à la mauvaise échelle. typiquement nos techniques pour générer des données synthétiques en exp mixture et connexification sont de ce type, et pour cela nous ne les explorons pas mais utilisons comme générateur de données uniquement.** Enfin, une approche originale et intéressante à la croissance des réseaux sont les réseaux biologiques. Ils appartiennent au champ de l'ingénierie morphogénétique dont DOURSAT est un pionnier, qui vise à concevoir des systèmes com-

plexes artificiels inspirés de systèmes complexes naturels et sur lesquels un contrôle des propriétés émergentes est possible [DOURSAT, SAYAMA et MICHEL, 2012]. Les *Machines Physarum*, qui sont des modèles d'une moisissure auto-organisée (*slime mould*) ont été prouvés comme résolvant de manière efficiente et par le bas des problèmes computationnellement lourds comme des problème de routage [TERO, KOBAYASHI et NAKAGAKI, 2006] ou des problèmes de navigation NP-complets comme le Problème du Voyageur de Commerce [ZHU et al., 2013a]. **C : (Florent) cela n'est pas de première importance je pense**
 Ils produisent des réseaux ayant des propriétés de coût-robustesse Pareto-efficiences [TERO et al., 2010], **C : (Florent) et alors, est ce que cela correspond à une réalité empirique?** repartir des trois sphères Muller Livet Sanders peut aider (empirique, conceptuel, du modèle) relativement proches en forme de réseaux réels (sous certaines conditions, voir [ADAMATZKY et JONES, 2010]). Ce type de modèles peut être d'intérêt dans notre cas puisque les processus d'auto-renforcement basés sur les flots sont analogues aux mécanismes de renforcement de lien en économie des transports.

C : comparison with Francois model for french railway (nothing published yet) - C Mieur (thèse soutenue?)

C : [LEVINSON, XIE et OCA, 2012] mécanismes induisant la croissance du réseau, gouvernance et économiques, très détaillés, basé sur enquêtes quali et modèle stats fittés sur vraies données

C : [XIE et LEVINSON, 2009b] compares centralized vs decentralized network growth

C : [LEVINSON et KARAMALAPUTI, 2003] fits statistical models, including multinomial logit, to find driver of highway network growth (on Twin Cities). Basic variables (length, change in accessibility) have expected behavior; there is a difference between interstate and local investments : local road growth is not affected by cost. Corresponds to requirement of equity in local territorial accessibility ?

C : [CHEN et LEVINSON, 2006] : la simulation comme outil pour apprendre aux élèves ingénieurs. Intéressant à utiliser pour l'aspect performatif, feedback des modèles sur les situations réelles / illustration des différents objectifs de chaque domaine : pourquoi et comment c'est intéressant de prendre en compte certains aspects selon les objectifs / perspectivisme appliqué : faire ce projet , l'évoquer ici.

C : [MIMEUR, 2016] la thèse de Mieur est un pont intéressant entre géographie et approches éco de Levinson (modèle de croissance type slime mould?). plus fait des stats spatiales pour lier croissance pop et accessibilité : checker si même résultats quand fera spatio-temp causalités sur réseau ferré et autoroutier et croissance pop. remarque : trucs bizarres, essaie d'expliquer pour petites villes, mais pas approprié, pb du choix de l'échelle, de ce qui est du bruit et du signal - semble tout mélanger : importance du preprocessing et traitement

du signal (cf correlations des taux de croissance). Tester effets fixes régions/départements ? fait GWR finalement ?

2.1.3 Modéliser la co-évolution

Modélisation Hybride

Les modèles de simulation qui incluent un couplage des dynamiques de la croissance urbaine et du réseau de transport sont relativement rares, et pour la plupart au stade de modèles stylisés. Une généralisation du modèle d'optimisation locale géométrique décrit précédemment a été développé dans [BARTHÉLEMY et FLAMMINI, 2009]. Comme pour le modèle de croissance de réseau routier dont il est l'extension, les mécanismes locaux n'ont pas de justification théorique ou thématique, et le modèle n'est de plus pas exploré et aucune connaissance géographique ne peut en être tirée. [LEVINSON, XIE et ZHU, 2007] prend une approche économique plus intéressante du point de vue des processus de développement de réseau impliqués, similaire à un modèle à quatre étapes (génération de flux origine-destination basés sur la gravité, attribution du traffic par Equilibre Utilisateur Stochastique) qui inclut coût de transport et congestion, couplé avec un module d'investissement routier qui simule les revenus des péages pour les agents qui construisent, et un module d'évolution d'usage du sol qui met à jour les actifs et emplois par modélisation de choix discrets. Les expériences montrent que l'usage du sol et le réseau en co-évolution mène à des retroactions positives renforçant les hiérarchies, mais sont loin d'être satisfaisantes pour deux raisons : d'une part la topologie du réseau n'évolue pas à proprement parler puisque seules les capacités et les flux changent dans le réseau, ce qui signifie que des mécanismes plus complexes sur de plus longues échelles de temps ne sont pas pris en compte, et d'autre part les conclusions sont assez limitées puisque le comportement du modèle n'est pas connu, les analyses de sensibilité étant faites sur un petit nombre d'espaces unidimensionnels : les mécanismes exhaustifs restent ainsi inconnus comme seuls des cas particuliers sont donnés dans l'analyse de sensibilité. D'un autre point de vue, [LEVINSON et CHEN, 2005] est aussi présenté comme un modèle de co-évolution mais correspond plus à une analyse statistique couplée puisqu'elle repose sur un modèle prédictif à chaîne de Markov. [RUI et BAN, 2011] décrit un modèle dans lequel le couplage entre usage du sol et la topologie du réseau est fait par un paradigme faible, l'usage du sol et l'accessibilité n'ayant pas de retroaction sur la topologie du réseau. [ACHIBET et al., 2014] décrit un modèle de co-évolution à une très petite échelle (échelle du bâtiment), dans lequel l'évolution du réseau et des bâtiments sont tous les deux régis par un agent commun (qui est influencé différemment par la topologie du réseau et la densité de population) ce qui implique une simplification trop grande des processus sous-jacents. En-

fin, un modèle hybride simple exploré et appliqué à un exemple jouet de planification dans [RAIMBAULT, BANOS et DOURSAT, 2014], repose sur les mécanismes d'accès aux activités urbaines pour la croissance des établissements avec un réseau s'adaptant à la forme urbaine. Les règles pour la croissance du réseau sont trop simples pour capturer les processus qui nous intéressent, mais le modèle produit à une petite échelle une large gamme de formes urbaines qui reproduisent les motifs typiques des établissements humains. A une échelle macroscopique et plus proche de la modélisation de système urbains que nous développerons dans la section suivante, [BAPTISTE, 1999] propose de coupler le modèle de croissance urbaine basé sur les migrations (introduit par l'application de la synergétique au système de ville par SANDERS dans [SANDERS, 1992]) avec un mécanisme d'auto-renforcement pour le réseau routier sans modification topologique (retroaction positive par seuils du différentiel flux-capacité sur la capacité). Guère de conclusions générales ne peuvent cependant être tirées de ce travail, autre que ce couplage permet de faire émerger une configuration hiérarchique (mais on sait par ailleurs que des modèles plus simples, un attachement préférentiel uniquement par exemple, permettent de reproduire ce fait stylisé) et que l'ajout du réseau produit un espace moins hiérarchique, permettant à des villes moyennes de bénéficier de la rétroaction du réseau de transport.

C : (Florent) pas assez de prise de hauteur sur cette partie pour une fois, on ne voit pas le tableau d'ensemble

C : (Florent) : cf renvoi remarques générales sur chapitre 1

C : (Juste) [BAPTISTE, 2010] : paper, quite the same as in thesis by Baptiste. [BADARIOTTI, BANOS et MORENO, 2007; MORENO, BADARIOTTI et BANOS, 2012] : remus and raumulus, inspiration for rbd model. relire thèse Moreno.

C (JR) : [BLUMENFELD-LIEBERTHAL et PORTUGALI, 2010] : hybrid model (largely discussed by Clara); network growth induces migration; would be interesting to test its abilities to produce various causality regimes (note : may be one indicator of how a model captures co-evolution?)

Modélisation de Systèmes Urbains

Une approche relativement proche des précédentes, mais ayant des caractéristiques propres, est celle de la modélisation intégrée des systèmes de villes. Dans la continuité des modèles Simpop pour modéliser les systèmes de villes, SCHMITT décrit dans [SCHMITT, 2014] le modèle SimpopNet qui vise à précisément intégrer les processus de co-évolution dans les systèmes de villes à longue échelle temporelle, typiquement par des règles pour un développement hiérarchique du réseau comme fonction des dynamiques des villes, couplées à celles-ci qui dépendent de la topologie du réseau. Malheureusement le modèle n'a pas été exploré ni étudié de manière plus approfondie,

et de plus est resté au niveau de modèle jouet. COTTINEAU propose une croissance endogène des réseaux de transport comme la dernière brique de construction de ses productions Marius [COTTINEAU, 2014] mais cela reste à un niveau conceptuel puisque cette brique n'a pas encore été spécifiée ni implémentée. Il n'existe à notre connaissance pas de modèle empirique ou appliqué à un cas concret se basant sur une approche de la co-évolution par les systèmes urbains vus par la Théorie Evolutive des Villes. Nous nous positionnerons particulièrement dans cette lignée de recherche dans cette thèse, vu l'importance que prendra la Théorie Evolutive dans notre démarche Théorique et de Modélisation comme nous le détaillerons par la suite. L'ensemble des briques est nécessaire pour comprendre les implications de ce positionnement, mais le lecteur pressé pourra directement consulter le chapitre 9 pour une synthèse des implications théoriques à différents niveaux d'abstraction. Typiquement, les hypothèses épistémologiques fondamentales tel le rôle des relations et de la configuration spatiales, ou la présence d'un équilibre - nous considérons les systèmes urbains comme des systèmes complexes adaptatifs, auto-organisés loin de l'équilibre, sont typiques de cette approche si on les considère conjointement. On voit bien l'opposition aux principes épistémologiques de l'économie géographique : [FUJITA, KRUGMAN et MORI, 1999] introduit par exemple un modèle évolutionnaire capable de reproduire une hiérarchie urbaine et une organisation typique de la Théorie des Places Centrales, mais repose toujours sur la notion d'équilibres successifs, et surtout considère un modèle "à-la-krugman" c'est à dire un espace à une dimension homogène. Cette approche peut être instructive sur les processus économiques en eux-mêmes mais aucunement sur les processus géographiques, qui incluent le déroulement des processus économiques dans l'espace géographique dans lequel les particularités sont essentielles. Notre travail s'attellera à montrer dans quelle mesure cette structure de l'espace peut être importante et également explicative, puisque les réseaux , et encore plus les réseaux physiques induisent des processus dépendants au chemin spatio-temporel et donc sensibles au singularités locales et propices aux bifurcations induites par la combinaison de celles-ci et de processus à d'autres échelles (par exemple la centralité induisant un flux).

Co-évolution

C : (Florent) constat à livrer d'emblée (pas de modèle de co-evolution dans algoSR) A1 : détailler ici que en effet par vraiment en notre sens.

C (JR) : une première entrée simple sur la co-evolution : couplage fort pour l'instant - sinon cf théorie (d'ailleurs y revenir sur multiples causality regimes : more links theory - reste) - [PAULUS, 2004] : evidence of co-evolution phenomena -> put when introduce co-evol. ;

développer “les lacunes à combler” : modèles fortement couplés plus ou moins multi-processes et multi-échelles ? (dans le temps au moins) - ce que nos modèles apportent - dans une vision de théorie intégrative.

2.2 UNE APPROCHE EPISTÉMOLOGIQUE

Un corolaire de la matière thématique introduite en chapitre 1 est le besoin d'une compréhension des disciplines impliquées elles-même pour être en mesure de construire des modèles hétérogènes intégrés. Les possibilités de couplage et d'intégration sont hautement déterminées par les approches existantes et les lacunes correspondantes qui ont été exposées dans la section précédente 2.1. Cela implique une étude épistémologique avancée dans chaque champ, que nous proposons de mener de manière quantitative et systématique. Ce choix délibéré pourrait occulter des considérations épistémologiques élaborées mais suit notre objectif d'investigations préliminaires pour la construction de modèles, en révélant potentiellement des directions de recherche.

Nous décrivons et explorons d'abord un algorithme de revue systématique algorithmique, qui reconstruit des corpus de références par une extraction sémantique itérative. Nous procédons ensuite à une analyse de réseaux, couplant réseau de citation et réseau sémantique, pour préciser les contours des disciplines impliquées. Nous suggérons finalement des possibles extensions vers de l'apprentissage non-supervisé et la fouille de texte complets pour une extraction automatique de la structure de modèles par exemple.

2.2.1 Revue Systématique Algorithmique

Une étude bibliographique étendue suggère une rareté des modèles quantitatifs de simulation qui intègrent à la fois la croissance urbaine et la croissance des réseaux. Cette absence pourrait être due aux intérêts divergents des disciplines concernées qui induiraient un manque de communication. Nous proposons de procéder à une revue de la littérature systématique et algorithmique pour donner des éléments de réponse quantitatifs à cette question. Un algorithme itératif formel pour construire des corpus de références à partir de mots-clés initiaux, basé sur l'analyse textuelle, est développé et mis en oeuvre. Nous étudions ses propriétés de convergence et procédons à une analyse de sensibilité. Nous l'appliquons ensuite à des requêtes représentatives de notre question spécifique, pour lesquelles les résultats tendent à confirmer l'hypothèse d'isolation des disciplines.

En recherche de modèles de co-évolution

Les réseaux de transport et l'usage du sol urbain sont connus pour être des composantes fortement couplées des systèmes urbains à différentes échelles [BRETAGNOLLE, PUMAIN et VACCHIANI-MARCUZZO, 2009]. **C : (Florent) c'est une affirmation très forte : pourquoi fermer le débat à ce stade ?** Une approche commune est de les considérer comme étant en co-évolution, tout en évitant les interprétations

trompeuses comme le mythe des effets structurants des infrastructures de transport [OFFNER, 1993]. Une question qui se présente rapidement est l'existence de modèles endogénisant cette co-évolution, i.e. prenant en compte simultanément la croissance urbaine et celle du réseau. Nous essayons d'y répondre par une revue systématique algorithmique. Nous proposons dans cette section, après un état de l'art rapide de la littérature existante, de développer cette approche en formalisant l'algorithme, dont les résultats sont ensuite présentés et discutés.

Modéliser les Interactions entre croissance urbaine et croissance des réseaux

Nous avons revu selon divers point de vue les efforts de modélisation des interactions entre territoires et réseaux dans la section précédente 2.1.

Analyse Bibliométrique

Avec l'avènement des nouveaux moyens techniques et des nouvelles sources de données, la revue de littérature classique tend à se coupler à des revues automatiques. Des techniques de revue systématique ont été développées, des revues qualitatives aux meta-analyses quantitatives qui permettent de produire des nouveaux résultats par combinaison d'études existantes [RUCKER, 2012]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'information qui par l'accès plus aisément à l'information rend difficilement justifiable l'omission de références clés [LISSACK, 2013]. Nous proposons de tirer parti de telles techniques pour traiter notre problème. En effet, l'observation de la bibliographie obtenue dans la section précédente soulève une hypothèse. On peut postuler sans risques à partir de la revue précédente 2.1 Il semble clair que toutes les briques sont présentes pour l'existence de modèles co-évolutifs mais des questionnements et objectifs différents semblent la stopper. Comme montré par [COMMENGES, 2013b] pour le concept de mobilité, pour lequel un "petit monde d'acteurs" relativement fermé, en l'occurrence les corsards des Ponts, a inventé une notion ad hoc, utilisant des modèles sans connaissance préalable d'un contexte scientifique plus général. On pourrait se trouver dans un cas similaire pour le type de modèles auxquels on s'intéresse. Des interactions restreintes entre des champs scientifiques travaillant sur les mêmes objets mais avec des objectifs et contextes divergents, et à des échelles différentes, pourrait être à l'origine de l'absence de modèles co-évolutifs. Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citation [NEWMAN, 2013] ou les réseaux de co-auteurs [SARIGÖL et al., 2014], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [CHAVALARIAS et COINTET,

2013], qui obtient une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. **C : (Florent) tu ne dis pas clairement pourquoi ses méthodes classiques ne suffisent pas** La méthode est particulièrement adaptée pour notre étude puisque nous voulons comprendre la structure du contenu des recherches sur le sujet. Nous appliquons une approche algorithmique décrite par la suite. L'algorithme procède par itérations pour obtenir un corpus stabilisé à partir de mots-clés initiaux, reconstruisant l'horizon sémantique scientifique autour d'un sujet donné.

DESCRIPTION DE L'ALGORITHME Soit A un alphabet, **C : (Florent) dans quelle théorie mathématique ? A1 : just a set of symbols** A^* les mots correspondants et $T = \cup_{k \in \mathbb{N}} A^{*^k}$ les textes de longueur finie sur celui-ci. Ce qu'on nomme une référence est pour l'algorithme un enregistrement avec des champs textuels représentant le titre, le résumé et les mots-clés. L'ensemble de références à l'itération n sera noté $\mathcal{C} \subset T^3$. **C : (Florent) pourquoi ? A1 : car titre, résumé, mots clés - on aggere dans tous les cas pour l'extraction** Nous supposons l'existence d'un ensemble de mots-clés \mathcal{K}_n , les mots-clés initiaux étant \mathcal{K}_0 . **C : (Florent) si j'ai bien compris, c'est l'utilisateur qui rentre \mathcal{K}_0 , c'est commode non ? A1 : c'est plutôt dans l'esprit de cet algo là. on pourrait aussi partir d'un corpus; mais ça c'est plutôt la partie suivante avec le réseau de citations.** Une itération procède de la manière suivante :

1. Un corpus intermédiaire brut \mathcal{R}_n est obtenu par une requête à un catalogue **C : (Florent) notamment tu dois avoir énormément de bruit, comment le gères tu ?** auquel on fourni les mots-clés précédents \mathcal{K}_{n-1} .
2. Le corpus total est actualisé par $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$. **C : (Florent) pas clair que sont \mathcal{C}_0 et \mathcal{K}_0**
3. Les nouveaux mot-clés \mathcal{K}_n sont extraits du corpus par Traitement du Langage Naturel (NLP), étant donné un paramètre N_k fixant le nombre de mot-clés. **C : (Florent) variable à chaque pas ? A1 : non**

L'algorithme termine quand la taille du corpus devient stable ou quand un nombre maximal d'itérations défini par l'utilisateur est atteint. La figure 1 synthétise le processus général.

RÉSULTATS De par l'hétérogénéité des opérations requises par l'algorithme (organisation des références, requêtes au catalogue, analyse textuelle), le langage Java s'est présenté comme une alternative raisonnable. Le code source est disponible sur le dépôt ouvert du projet **C : (Florent) lien ?**¹. Les requêtes au catalogue, qui consistent à

¹ à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR>

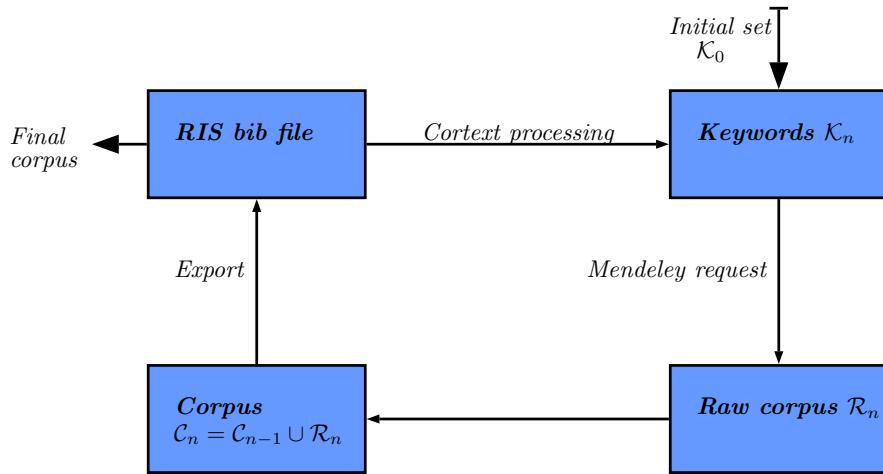


FIGURE 1 : Architecture globale de l'algorithme, incluant des détails d'implémentation : la requête au catalogue est faite via l'API Mendeley ; les corpus finaux sont sous forme de fichiers RIS.

récupérer un ensemble de références à partir d'un ensemble de mots-clés, sont faites via l'API du logiciel Mendeley [MENDELEY, 2015] qui permet un accès ouvert à une base de données conséquente. L'extraction des mots-clés est effectuée par techniques d'Analyse Textuelle (NLP) selon le processus donné dans [CHAVALARIAS et COINTET, 2013], via un script Python qui utilise [BIRD, 2006].

Une preuve formelle de convergence de l'algorithme n'est guère envisageable puisque qu'elle dépendra de la structure empirique inconnue des résultats de requête et d'extraction de mots-clés. Il est donc nécessaire d'étudier le comportement de l'algorithme de manière empirique. Comme présenté en figure 2, l'algorithme a de bonnes propriétés de convergence mais diverse sensibilités à N_k . Nous étudions également la cohérence lexicale interne des corpus finaux et fonction du nombre de mots-clés. Comme attendu, des valeurs faibles produisent des corpus plus cohérents, mais la variabilité lorsque qu'elles augmentent reste raisonnable.

Lorsque l'algorithme a été partiellement validé, **C : (Florent)** avec quels mots clés as tu validé empiriquement la convergence de l'algo ? on peut l'appliquer à notre question. Nous partons de cinq différentes requêtes initiales qui ont été manuellement extraites des divers domaines identifiés dans la bibliographie (qui sont “city system network”, “land use transport interaction”, “network urban modeling”, “population density transport”, “transportation network urban growth”). **C : (Florent) pourquoi ce mots là (par ex pas coevolution?)** Nous prenons l'hypothèse la plus faible pour le paramètre $N_k = 100$, **C : (Florent) pourquoi est ce weak?** au sens où les domaines atteints devraient être moins restreints. Après avoir construit les corpus, nous étudions leur cohérence lexicale comme un

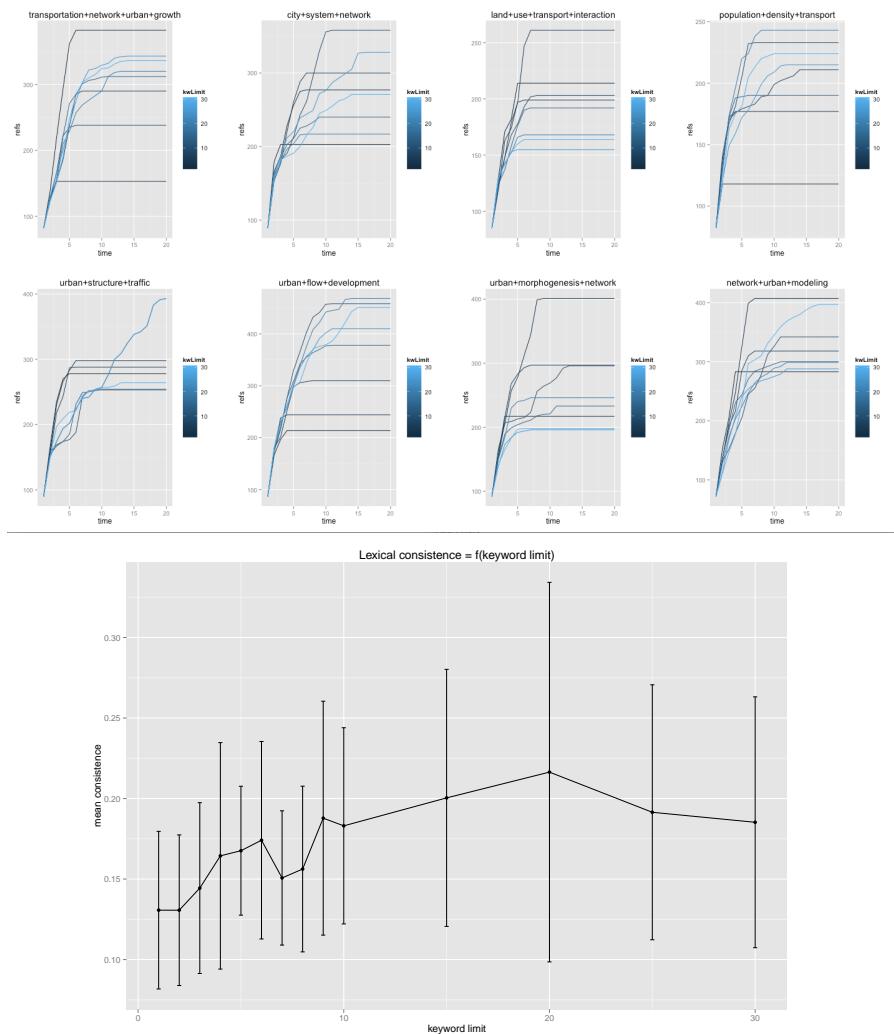


FIGURE 2 : C : (Florent) illisible

Corpus	1	2	3	4	5
1 ($W=3789$)	1	0	0.0719	0.0078	0.0724
2 ($W=5180$)	0	1	0.0338	0	0.0125
3 ($W=3757$)	0.0719	0.0338	1	0.0100	0.1729
4 ($W=3551$)	0.0078	0	0.0100	1	0.0333
5 ($W=8338$)	0.0724	0.0125	0.1729	0.0333	1

TABLE 1 : Matrice symétrique des proximités lexicales entre les corpus finaux, définies comme la somme des co-occurrences totale de mots-clés finaux entre corpus, normalisé par le nombre de mots-clés finaux (100). Les valeurs obtenues sont considérablement faibles, ce qui confirme que les corpus sont éloignés de manière significative. La taille des corpus finaux est donnée par W . **C : (Arnaud) a développer**

indicateur de réponse à notre question initiale. De grande distances devraient confirmer l'hypothèse formulée ci-dessus, i.e. que des disciplines auto-centrées pourraient être à l'origine d'un manque d'intérêt pour des modèles co-évolutifs. La table 1 montre les valeurs de la proximité lexicale relative, qui est significativement basse, **C : (Florent) comment peut on en juger ? A1 : c'est un des flaws, on n'a pas de null model..** confirmant notre hypothèse.

Les développements possibles incluent la construction de réseaux de citation via un accès automatique à Google Scholar qui fournit les citations entrantes. La confrontation des coefficients inter-clusters pour le réseau de citations entre les différents corpus avec la cohérence lexicale est un aspect clé d'une validation approfondie des résultats.

L'absence peu explicable a priori de modèles qui simulent la co-évolution des réseaux de transport et de l'usage du sol urbain, qui se confirme à première vue par un état de l'art couvrant des domaines disparates, pourrait être due à l'absence de communication entre les disciplines scientifiques étudiant différents aspects du problème. Nous avons proposé une méthode algorithmique pour donner des éléments de réponse par l'extraction de corpus basée sur l'analyse textuelle. Les premiers résultats numériques semblent confirmer l'hypothèse. Cependant, une telle analyse quantitative ne doit pas être considérée seule, mais devrait plutôt venir comme soutien à des études qualitatives qui peuvent être l'objet de développements futurs, comme celle menée dans [COMMENGES, 2013b], dans laquelle des questionnaires avec des acteurs historiques fournissent des informations extrêmement pertinentes. **C : (Florent) tu l'as déjà dit; il y a d'autres arguments à mobiliser (comme les cas d'application potentiels de tes**

modèles) - sur des temporalités si longues qu'ils ne bénéficient pas ou peu de financements ad hoc

2.2.2 *Bibliométrie Indirecte par Analyse de Réseaux Complexes*

Comme décrit précédemment, l'analyse sémantique des corpus finaux ne contient pas la totalité de l'information sur les liens entre disciplines ni sur les motifs de propagation de la connaissance scientifique comme ceux contenus dans les réseaux de citations par exemple. De plus, la collection des données dans l'algorithme précédent est sujette à convergence vers des thèmes relativement auto-cohérents de par la structure propre de la méthode. On pourrait obtenir plus d'information sur les motifs sociaux de choix ontologiques pour la modélisation en étudiant les communautés dans des réseaux plus larges, ce qui correspondrait plus à des disciplines (ou des sous-disciplines selon le niveau de granularité). Nous proposons de reconstruire les disciplines autour de notre thématique, pour obtenir une vue plus précise de l'interdisciplinarité et du paysage scientifique sur notre sujet.

2.2.3 *Discussion*

Vers une modélisation des thèmes et une extraction automatique du contexte

A possible direction to strengthen our quantitative epistemological analysis would be to work on full textes related to the modeling of interaction between networks and territories, with the aim to automatically extract thematics within articles. The idea would be to perform some kind of automatized modelography, with possible features to be extracted that would be ontologies, model architecture or structures, scales, or even typical parameter values. It is not clear to what degree structure of models can be extracted from their description in papers and it surely depends on the discipline considered. For example in a framed field such as transportation planning, using a pre-defined ontology (in the sense of dictionary) and a fuzzy grammar could be efficient to extract information as the discipline is relatively formatted. In theoretical and quantitative geography, beyond the barrier of language, information organisation is surely less subject to unsupervised data-mining because of the more literary nature of the discipline : synonyms and figures of speech are generally the norm in good level human sciences writing, fuzzing a possible generic structure of knowledge description.

Depending on extended results of the two previous sections and on thematic requirements (huge need of knowledge on precise models structure, that may appear when trying to construct more specialized

operational models), this project may be conducted with more or less investment.

Réflexivité

The methodology developed here is particularly interesting since it is reflexive, i.e. it can be used on our work itself. Therefore, an other application will be the reflexivity of our thesis : we attend to proceed to similar analysis on our proper bibliography (and possibly its evolution, available via git history), to understand our patterns of knowledge, possible gaps or unveil unexpected developments. The detailed development is done in Appendix ??.

2.3 REVUE SYSTÉMATIQUE ET MODÉLOGRAPHIE

C (JR) : la modélographie doit logiquement arriver après les études d'épistemo quanti, qui ont permis de donner un aperçu de l'horizon scientifique

An ongoing work is the production of a synthesis of this overview, from a modular modeling point of view, combined with a purpose and scale classification. Already mentioned, modular modeling consists in the integration of heterogeneous processes and implementation of processes in order to extract the set of mechanisms giving the best fit to empirical data [COTTINEAU, CHAPRON et REUILLOON, 2015]. We can thus classify models described here according to their building bricks in terms of processes implemented and thus identify possible coupling potentialities. This work is a preliminary step for the analysis in quantitative epistemology developed in chapter ??.

2.3.1 Revue systématique et Meta-analyse

Tandis que les études menées précédemment proposaient de construire un horizon global de l'organisation des disciplines s'intéressant à notre question, nous proposons à présent une étude plus ciblée des caractéristiques de modèles existants. Nous proposons pour cela dans un premier temps une revue systématique, c'est à dire la construction d'un corpus répondant à certaines contraintes, suivie d'une meta-analyse, c'est à dire une tentative d'explication de certaines caractéristiques des modèles par des modèles statistiques.

C (JR) : également tenter une classif endogène des modèles : selon les caractéristiques récupérées.

2.3.2 Modélographie

Nous passons à présent à une analyse mixte inspirée par les résultats précédents, notamment pour la classification. Elle a pour but d'extraire et de décomposer précisément les ontologies, échelles et processus, puis d'étudier des liens possibles entre ces caractéristiques des modèles et le contexte dans lequel ils ont été introduits. Il s'agit ainsi de la meta-analyse en quelque sorte, que nous désignerons ici par modélographie.

caractéristiques : temporal and spatial span, scales, équilibre ?,

TODO : "bon choix" de carac implique une bonne modularité de la classif obtenue dans une classif a priori (car on veut discriminer bien les modèles qu'on connaît et qu'on juge différent) → faire une random forest regression et regarder structure endogène; comparer à régressions simples.

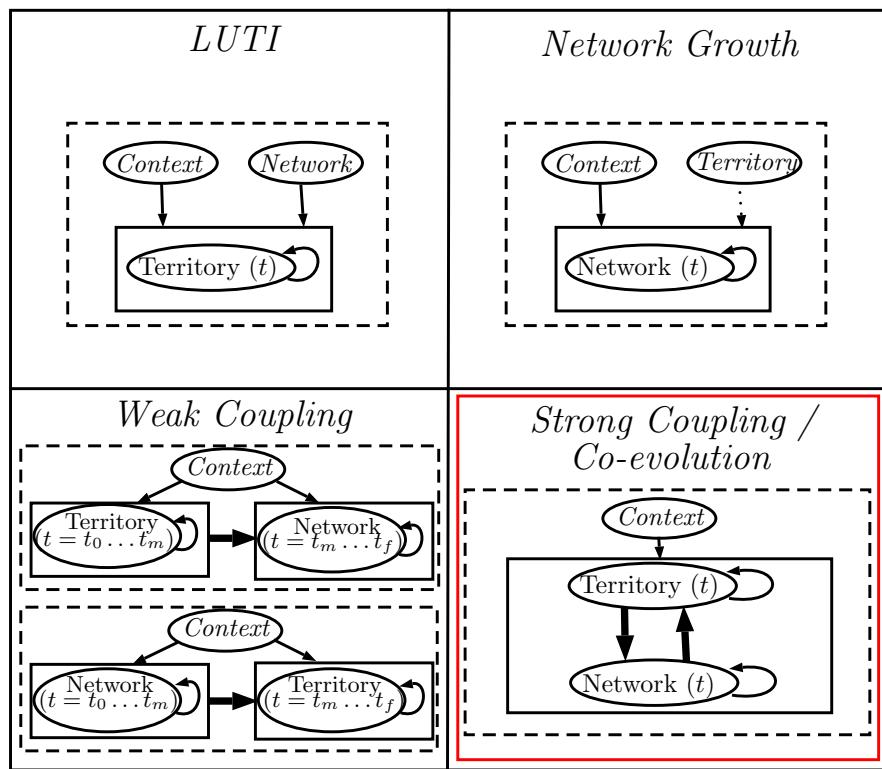


FIGURE 3 : Représentation schématique de la distinction entre différents types de modèles couplant territoires et réseaux.

2.3.3 *Discussion*

CONCLUSION DU CHAPITRE

3

POSITIONNEMENTS

Toute activité de recherche serait, selon certains observateurs, nécessairement politisée, de par pour commencer le choix de ses objets. Ainsi, RI POLL alerte contre l'illusion d'une recherche objective et les dangers de la technocratie [RI POLL, 2017]. Nous ne rentrerons pas dans ces débats bien trop vastes pour être traités même en un chapitre, puisqu'il rejoignent des thèmes de sciences politiques, d'éthique, de philosophie, liés par exemple à la gouvernance scientifique, à l'insertion de la science dans la société, à la responsabilité scientifique. Il est clair que même des sujets *a priori* intrinsèquement objectifs, comme la physique des particules et des hautes énergies, ont des implications regardant d'une part les choix de leur financements et les externalités associées (par exemple, l'existence du CERN a largement contribué au développement du calcul distribué), mais d'autre part aussi les applications potentielles des découvertes qui peuvent avoir des répercussions sociales considérables. En biologie, l'éthique est au cœur des principes fondateurs des disciplines, comme en témoignent les débats soulevés par l'émergence de la biologie synthétique [GUTMANN, 2011]. Les tenants d'approche prudentes dans celle-ci se recoupent avec la biologie intégrative, or les Sciences Intégratives défendues par PAUL BOURGINE, mises en oeuvre par l'intermédiaire du campus digital Unesco CS-DC¹, ont typiquement la responsabilité sociale et l'implication citoyenne au cœur de leur cercle vertueux. En sciences humaines, comme les recherches interagissent avec les objets étudiés (en quelque sorte l'idée des *interactive kind* de HACKING [HACKING, 1999]), les implications politiques et sociales de la recherche sont bien évidemment indiscutables. Là où il y aurait matière à discussion, et nous y reviendrons en ouverture 9.3.3 car il s'agira d'une des questions ouvertes posées par notre recherche et sa démarche dans leur ensemble, serait sur la compatibilité des méthodes systématisques et *evidence-based* avec les sciences sociales, autrement dit dans quelle mesure peut-on s'extraire de certains dogmatismes encore plus marqués lors de l'usage de théorie politiques². Nous resterons ici à un niveau épistémologique, c'est à dire à des réflexions sur la nature et le contenu des connaissances scientifiques au sens large, c'est à dire co-construites et validées au sein d'une communauté imposant certains critères de scientificité, bien sûr évolutifs puisque nous nous posi-

¹ <https://www.cs-dc.org/>

² MONOD montre par exemple les désastres liés aux "niaiseries épistémologiques" déroulant de l'application littérale de la dialectique matérialiste marxiste à l'épistémologie du vivant.

tionnerons pour la systématisation de certains. Mais donc, même en restant à ce niveau, des prises de positions sont nécessaires, celles-ci pouvant être épistémologiques, méthodologiques, thématiques. Ces dernières ont déjà été ébauchée dans les deux chapitres précédents par les choix des objets d'étude, des problématiques, et seront renforcées à mesure de la progression pour finalement être synthétisées en Chapitre 9. Nous proposons ici un exercice relativement original mais que nous jugeons nécessaire pour une lecture plus fluide de la suite, qui consiste en le développement précis de certains positionnements qui ont une influence particulière dans notre démarche de recherche. Par exemple, le travail en données quasi-intégralement ouverte et en architecture modulaire résulte de notre exigence de reproductibilité. L'utilisation des modèles et la manière de les explorer de notre vision du calcul intensif. Dans une première section (3.1), nous développons des exemples pour illustrer le besoin et la difficulté de reproductibilité, ainsi que les liens avec des nouveaux outils pouvant la favoriser mais aussi la mettre en danger. Dans une deuxième section (3.2), nous argumentons sous forme d'essai pour un usage raisonné des données massives et du calcul intensif, et illustrons notre positionnement par rapport à l'exploration des modèles par une étude de cas méthodologique pour l'exploration de la sensibilité des modèles aux conditions initiales. Enfin, la dernière section (3.3) explicite modestement des positions épistémologiques, notamment concernant le courant dans lequel nous nous plaçons, la complexité des objets en sciences sociales, et la nature de la complexité de manière générale. Le lecteur très familier avec les commandements de BANOS [BANOS, 2013] pourra éventuellement sauter les deux premières sections à part s'il est intéressé par des illustrations pratiques originales, notre positionnement étant très similaire et ne divergeant que sur des subtilités mineures pour les sujets évoqués dans ces sections.

* * *

*

Ce chapitre est composé de divers travaux. La première section est inédite. La deuxième section rend compte pour sa première partie du contenu théorique de [RAIMBAULT, 2016c], et pour sa deuxième partie des idées présentées dans [COTTINEAU et al., 2017]. La troisième section reprend dans sa première partie les bases épistémologiques de [RAIMBAULT, 2017g] approfondies par [RAIMBAULT, 2017a], est inédite pour sa deuxième partie et rend compte de [RAIMBAULT, 2017c] pour sa dernière partie.

3.1 REPRODUCIBILITÉ

La force de la Science vient de la nature cumulative et collective de la recherche, puisque les progrès sont faits lorsque, comme NEWTON l'a bien posé, on "se tient sur les épaules de géants", au sens que l'entreprise scientifique à un temps donné repose sur l'ensemble du travail précédent et qu'aucune avancée ne serait possible sans construire dessus. Cela inclut le développement de nouvelles théories, mais aussi l'extension, le test et la falsification de précédentes : l'avancée dans la construction de la tour signifie aussi la déconstruction de certaines briques obsolètes. Cet aspect de validation par les pairs et de remise en question constante est aussi ce qui légitime la Science pour une connaissance plus robuste et un progrès sociétal basés sur une connaissance d'un univers objectif, par rapport aux systèmes dogmatiques qu'ils soient politiques ou religieux [BAIS, 2010].

La reproductibilité semble être de plus en plus pratiquée de manière effective [STODDEN, 2010] et les moyens techniques pour l'achever sont toujours plus développés (comme par exemple les outils pour déposer les données ouvertes, ou pour être transparent dans le processus de recherche comme git [RAM, 2013], ou pour intégrer la création de document et l'analyse de données comme knitr [XIE, 2013]), au moins dans le champ de la modélisation et de la simulation. Cependant le diable est bien dans les détails et des obstacles jugés dans un premier temps comme mineurs peuvent rapidement devenir un fardeau pour reproduire et utiliser des résultats obtenus dans des recherches précédentes. Nous décrivons deux études de cas où les modèles de simulation sont en apparence hautement reproductibles mais se révèlent vite des puzzles pour lesquels l'équilibre de temps de recherche passe rapidement sous zéro, au sens où essayer d'exploiter leur résultats coûtera plus en temps que de développer entièrement des modèles similaires.

3.1.1 *Explicitation, documentation et implémentation des modèles*

Sur le Besoin d'expliciter le modèle

Un mythe à la vie dure (auquel nous essayons en fait nous-même d'échapper) est que fournir le code source complet et les données seront une condition suffisante pour la reproductibilité, puisque la reproductibilité computationnelle complète implique un environnement similaire ce qui devient vite ardu à produire comme le montre [HATTON et WARR, 2016]. Pour résoudre ce problème, [HUNG et al., 2016] propose l'utilisation de conteneurs Dockers qui permet de reproduire même le comportement de logiciels avec interface graphique indépendamment de l'environnement. C'est d'ailleurs une des direction courantes de développement d'OpenMole, pour simplifier le packaging des bibliothèques et des modèles en binaire (cf. R. REUILLO

dans [RAIMBAULT, 2017d]). Dans tous les cas, la reproductibilité a des dimensions supplémentaires, il ne s'agit pas de l'objectif unique qui serait est de produire exactement les mêmes graphes et analyses statistiques, en supposant que le code fournit est celui qui a été effectivement utilisé pour produire les résultats donnés. Tout d'abord, doivent être autant que possible indépendants de l'implémentation (c'est à dire du langage, des bibliothèques, des choix de structures de données et de type de programmation) pour des motifs clairs de robustesse. Ensuite, en relation avec le point précédent, un des buts de la reproductibilité est la réutilisation des méthodes ou résultats comme base ou modules pour une recherche future (ce qui comprend une implémentation dans un autre langage ou une adaptation de la méthode), au sens que la reproductibilité n'est pas la possibilité stricte de répliquer car elle doit être adaptable [DRUMMOND, 2009].

Notre premier cas d'étude suit exactement ce schéma, puisqu'il a sans aucun doute été conçu pour être partagé avec la communauté et utilisé, s'agissant d'un modèle de simulation fourni avec la plate-forme de modélisation agent NetLogo [WILENSKY, 1999]. Le modèle est également disponible en ligne [DE LEON, FELSEN et WILENSKY, 2007] et est présenté comme un outil pour simuler les dynamiques socio-économiques des résidents à bas revenus d'une ville au sein d'un environnement urbain synthétique, généré pour ressembler en terme de faits stylisés à la ville réelle de Tijuana, Mexico. Globalement, le modèle fonctionne de la façon suivante : (i) à partir de centre urbains, une distribution d'usage du sol est générée par modélisation procédurale similaire à [LECHNER et al., 2006], c'est à dire des routes sont générées de proche en proche selon des règles géométriques et de hiérarchie locales, et un usage du sol ainsi qu'une valeur est attribué en fonction des caractéristique du patch (distance au centre, à la route) ; (ii) dans cet environnement urbain sont simulées des dynamiques résidentielles de migrants, qui cherchent à optimiser une fonction d'utilité dépendant du coût de la vie et de la configuration des autres migrants. A part fournir le code source, le modèle n'est que peu documenté dans la littérature ou dans les commentaires et la description de l'implémentation. Les commentaires qui suivent sont basés sur l'étude de la partie du modèle simulant la morphogenèse urbaine (setup pour la composante "dynamiques résidentielles") comme il s'agit de notre contexte global d'étude. Dans le cadre de cette étude, le code source a été modifié et commenté, dont la dernière version est disponible sur le dépôt du projet³.

FORMALISATION RIGOUREUSE Une partie évidente de la construction d'un modèle est sa formalisation rigoureuse dans un cadre formel distinct du code source. Il n'y a bien sûr aucun langage universel pour le formuler [BANOS, 2013], et de nombreuses possibilités sont

³ at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite>

offertes par de nombreux champs (e.g. UML, DEVS, formulation mathématique pure), mais l'étape de formalisation précise, qui suit généralement une description plus intuitive donnant les idées et processus dominants ("rationnelle"), ne peut pas être sautée. On pourrait se dire que le code source y est équivalent, mais ce n'est pas exactement vrai car on pourrait alors ne plus distinguer certains choix d'implémentation de la structure du modèle. Aucun article ni documentation n'accompagne le modèle ici, au delà de la documentation embarquée NetLogo, qui ne décrit que de manière thématique en langage naturel les idées derrière chaque étape sans plus développer et fournir de l'information sur le rôle des différents éléments de l'interface. Comme ces éléments manquent ici, le modèle n'est guère utilisable tel quel. On pourrait nous objecter ici que la partie que nous étudions est une procédure d'initialisation et non le cœur du modèle : nous maintenons que l'ensemble des procédures doit être également documenté et implémenté avec un soin équivalent, ou pointer vers une référence extérieure dans le cas d'utilisation d'un modèle tiers, comme nous le faisons d'ailleurs pour le couplage effectué en [3.2](#).

Une telle formulation est essentielle pour que le modèle soit compris, reproduit et adapté ; mais elle évite également des biais d'implémentation comme

- Des éléments architecturaux dangereux : dans le modèle, le contexte du monde est une sphère, ce qui n'est pas raisonnable pour un modèle à l'échelle d'une ville. Les agents peuvent "sauter" dans la représentation euclidienne, ce qui n'est pas acceptable pour une projection en deux dimensions du monde réel. Pour éviter cela, de nombreux tests et fonctions subtils sont utilisés, incluant des pratiques déconseillées (e.g. mort d'agents basée sur leur position pour les empêcher de sauter).
- Manque de cohérence interne : par exemple la variable de patch `land-value` utilisée pour représenter différentes quantités géographiques à différentes étapes du modèle (morphogenèse et dynamiques résidentielles), ce qui devient une inccohérence interne quand les deux étapes sont couplées lorsque l'option permettant de faire croître la ville est activée.
- Erreur de code : dans un langage non typé comme NetLogo, le mélange des types peut conduire à des erreurs inattendues à l'exécution, ou même des bugs non détectables directement et alors plus dangereux. C'est le cas de la variable de patch `transport` dans le modèle (même si aucune erreur ne survient dans la majorité des configurations depuis l'interface, ce qui est plus dangereux comme le développeur pense que l'implémentation est sûre). De tels problèmes devraient être évités si l'implémentation est faite à partir d'une description exacte du modèle.

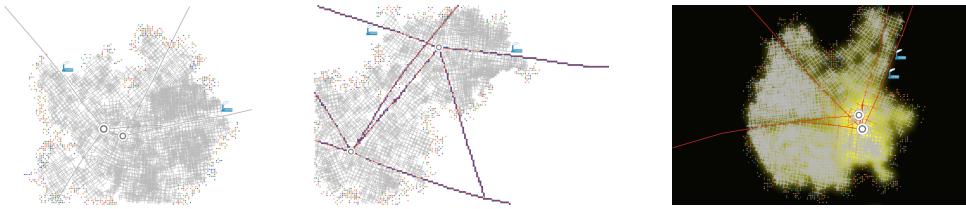


FIGURE 4 : Exemple d'amélioration simple dans la visualisation qui peut aider à apprêhender les mécanismes impliqués par le modèle. (Gauche) Exemple de sortie originale ; (Centre) Visualisation des routes principales (en rouge) et de l'attribution des patches sous-jacente, qui suggère de possibles biais d'implémentation dans l'utilisation de la trace discrete des routes pour garder trace de leur position ; (Droite) Visualisation des valeurs foncières en utilisant un gradient de couleur plus lisible. Cette étape confirme l'hypothèse, par la forme de la distribution des valeurs, que l'étape de morphogenèse est un détour non-nécessaire pour générer un champ aléatoire pour lequel des simples mécanismes de diffusion devrait fournir des résultats similaires, comme détaillé dans le paragraphe sur l'implémentation. Initialement, l'interface du modèle ne permet pas ces options de visualisation, ces à dire se limite à la première image. On ne peut se rendre compte des processus en jeu pour la morphogenèse, liés aux patches de route et au valeurs foncières se diffusant.

IMPLÉMENTATION TRANSPARENTE Une implémentation totalement transparente doit être attendue, incluant une certaine ergonomie dans l'architecture et le code, mais aussi dans l'interface et la description du comportement attendu du modèle.

COMPORTEMENT ATTENDU DU MODÈLE Quelle que soit la définition, un modèle ne peut pas être réduit à sa formulation et/ou implémentation, comme le comportement attendu ou l'utilisation du modèle peuvent être vu comme des parties du modèle lui-même. Dans le cadre du perspectivisme de GIERE [GIERE, 2010c], la définition du modèle inclut le motif de l'utilisation mais aussi l'agent qui vise à l'utiliser. Pour cela une explication minimale du comportement du modèle et une exploration du rôle des paramètres est fortement recommandé pour décroître les chances de mauvais usage ou mauvaises interprétations de celui-ci. Cela inclut des graphes simples obtenus immédiatement à l'exécution sur la plateforme NetLogo, mais aussi un calcul d'indicateurs pour évaluer les sorties du modèle. Il peut aussi s'agir de visualisations améliorées pendant l'exécution et l'exploration du modèle, comme le montre la figure 4.

Sur le besoin d'exactitude dans l'implémentation du modèle

Des divergences potentielles entre la description du modèle dans un article et les processus effectivement implémentés peuvent avoir des conséquences graves sur la reproductibilité finale. Le modèle de crois-

sance du réseau routier donné dans [BARTHÉLEMY et FLAMMINI, 2008] est un exemple d'une telle discrépance. Une implémentation stricte des mécanismes du modèle produit des résultats légèrement différents de ceux présentés dans le papier, et comme le code source n'est pas fourni nous devrions tester différentes hypothèses sur des mécanismes possibles ajoutés par le programmeur (qui semble être une règle de connexion aux intersections sous un certain seuil de distance). Des leçons qui peuvent éventuellement être tirées de cet exemple, qui rejoignent partiellement mais complètent celle tirées dans l'étude de cas précédente, sont

- la nécessité de fournir le code source
- la nécessité de fournir une description de l'architecture en même temps que le code (si la description du modèle est faite dans un langage trop loin de spécification architecturales) afin d'identifier des biais possibles d'implémentation
- la nécessité de procéder à des explorations explicites du modèle et de les détailler, ce qui dans ce cas aurait permis d'identifier de possibles biais d'implémentation.

Rendre le dernier point obligatoire pourrait assurer un risque limité de falsification puisqu'il est généralement plus compliqué de falsifier des résultats d'exploration plutôt que d'explorer effectivement le modèle. On pourrait imaginer une expérience pour tester le comportement général d'un sous-ensemble de la communauté scientifique au regard de la reproductibilité, qui consisterait en l'écriture d'un faux papier de modélisation dans l'esprit de [ZILSEL, 2015], dans lesquels des résultats opposés aux résultats effectifs d'un modèle donné seraient fournis, sans fournir l'implémentation du modèle. Un premier test serait de tester l'acceptation d'un papier clairement non reproductible dans divers journaux, si possible avec un contrôle sur les éléments textuels (par exemple en utilisant ou non des "buzzwords" chers au journal). Selon les résultats, une expérience plus poussée serait de fournir l'implémentation open source mais toujours avec des résultats modifiés plus ou moins fortement, afin de tester si les reviewers essayent effectivement de reproduire les résultats quand ils demandent le code (dans des capacités de calcul limitées bien sûr, le HPC n'étant pas encore largement disponibles en sciences sociales). Notre intuition est que les résultats obtenus seraient fortement négatifs, vu les difficultés rencontrées par une exigence de discipline de reproduction indépendante lors de nombreuses relectures, même pour des revues faisant de la reproductibilité une condition *sine qua non* de la publication, les auteurs trouvant des astuces pour se dérober aux contraintes (postuler que des données de simulation ne sont pas des données, ne fournir qu'une version agrégée inutile du jeu de données utilisées, etc. ; nous reviendrons sur le rôle des données plus loin).

3.1.2 Exploration interactive et production des résultats

L'usage d'applications interactives pour la fouille de données a des avantages non discutables, tel qu'une familiarisation avec la structure des données par une vue d'ensemble qui serait beaucoup plus laborieuse voire impossible autrement. C'est la même idée sous-jacente qui justifie l'interactivité pour l'exploration préliminaire des modèles basé-agent intégrée à des plateformes comme NetLogo [WILENSKY, 1999] ou Gamma [Cit. gamma]. C'était d'ailleurs un objectif couplé qu'avait initialement [REY-COYREHOURCQ, 2015], c'est à dire une intégration complète de l'exploration fine des modèles et de la production des graphes de sortie ainsi que leur exploration interactive. Comme le rappelle R. Reuillon (Entretien du 11/04/2017, voir D.5), la plateforme OpenMole qui devait accueillir cette couche supplémentaire était loin d'être mature à l'époque et ne l'est toujours pas aujourd'hui, puisque l'état de l'art de telles pratiques est en pleine construction et bouleversements réguliers [HOLZINGER, DEHMER et JURISICA, 2014]. Des difficultés au regard de la reproductibilité, qui nous concernent particulièrement ici, sont récurrentes et loin d'être résolues. En effet, il faut bien situer la position de ces outils et méthodes comme une aide cognitive préliminaire⁴, mais peu souvent comme permettant la production de résultats finaux : lorsque les paramètres ou dimension se multiplient, l'export d'un graphe est bien souvent déconnecté de l'information complète ayant conduit à sa production. De la même manière, l'utilisation de notebooks intégrés tel Jupyter, permettant d'intégrer analyses et rédaction du compte-rendu, peut devenir dangereux car on peut justement revenir sur un script, tester différentes valeurs d'un paramètre, et perdre les valeurs qui avaient produit un graphe donné. L'utilisation de versioning peut être une solution partielle mais souvent lourde. Dans l'idéal, tout logiciel interactif permettant l'export de résultats devrait en même temps exporter un script ou une description exacte et utilisable permettant d'arriver exactement à ce point à partir des données brutes. La plupart des applications d'exploration interactives de données spatio-temporelles sont à ce regard relativement immatures scientifiquement, car même dans le cas où elles sont totalement honnêtes et transparentes sur les analyses présentées à l'utilisateur, ce qui n'est malheureusement pas la règle, les tâtonnements d'exploration progressive ne sont pas reproductibles et la méthode d'extraction de caractéristiques est ainsi relativement aléatoire. En poussant le raisonnement, leur utilisation révélerait plutôt l'aveu d'une faiblesse d'un manque de méthodes systématiques accompagnant la découverte de motifs dans des données spatio-temporelles complexes de manière efficace. De manière très visionnaire, BANOS avait déjà mis en garde contre "les dangers de

⁴ que nous ne jugeons pas superficielle puisque nous les mobilisons au moins par deux fois par la suite, voir 5.1 et 5.2

la jungle” des données dans [BANOS, 2001], quand il souligne très justement que l’exploration interactive doit nécessairement se doubler d’indicateurs locaux adaptés, mais surtout d’outils d’exploration automatisés et de critère d’évaluation des choix faits et des motifs découverts par l’utilisateur. On revient encore à l’idée d’une plate-forme intégrée dont OpenMole pourrait être un précurseur. La combinaison des capacités cognitives humaines au traitement machine, notamment pour des problèmes de vision par ordinateur, ouvre des possibilités de découvertes inédites, encore plus via une utilisation collective comme en témoigne le Galaxy Zoo [RADDICK et al., 2010]. Les résultats d’un crowdsourcing de la cognition humaine peuvent rivaliser avec les techniques automatiques les plus avancées comme le montre [KOCHE et STISEN, 2017] pour l’exemple de la comparaison de cartes spatiales. Ces possibilités ne doivent cependant pas être sur-estimées ou utilisées à mauvais escient, et les questions d’intégration efficiente homme-machine sont d’ailleurs totalement ouvertes. Dans le domaine de la visualisation de l’information géographique, [PFAENDER, 2009] introduit une sémiologie spécifique visant à favoriser l’exploration de grands jeux de données hétérogènes, et l’expérimente sur une application spécifique : il s’agit d’une avancée considérable vers une plateforme intégrée et une exploration interactive saine et reproductible, les directions d’exploration répondant à des modèles basés sur les sciences cognitives.

3.1.3 Perspectives

Encore une fois, la reproductibilité et la transparence sont des éléments essentiels incontournables de la science contemporaine, liés aux pratiques de science ouverte et d’accès ouvert. Beaucoup d’exemples (voir un récent en économie expérimentale dans [CAMERER et al., 2016]) dans diverses disciplines montrent le manque de reproductibilité des résultats des expériences, alors que celle-ci doit pouvoir conduire à une falsification ou à une confirmation de ces résultats. La falsification est une pratique coûteuse car demandant un certain investissement au détriment de sa propre recherche [CHAVALARIAS et al., 2005]. Elle pourrait ainsi être rendue plus efficiente grâce à une transparence augmentée. Des outils spécialement dédiés à une reproductibilité directe, souvent permise par l’ouverture, devraient accroître la performance globale de la science. Mais l’accès ouvert a des impacts bien plus larges que la science elle-même : [TEPLITSKIY, LU et DUEDE, 2015] montre un transfert des connaissances scientifiques accru vers la société dans le cas d’articles ouverts, notamment par des intermédiaires comme Wikipedia.

Le développement et la systématisation de standards et de bonnes pratiques, de manière conjointe sur les différentes problématiques évoquées, est une condition nécessaire à une rigueur scientifique qui

devrait être uniforme au travers de l'ensemble des disciplines existantes. Nous construisons par exemple des exemples d'outils facilitant le flot de production scientifique, ceux-ci étant détaillés en Appendice E.3. Par exemple, pour les sciences computationnelles, on a déjà évoqué les potentialités de l'utilisation de git qui s'étendent en fait sans contrainte de disciplines ni de types de recherche si les bonnes adaptations sont introduites. Le suivi précis de l'ensemble des étapes d'un projet, gardé en historique offrant la possibilité de revenir à n'importe laquelle à tout moment, mais aussi de travailler de façon collaborative, plus ou moins parallèlement selon les besoins en utilisant les branches, est un exemple de service fourni par cet outil. Un exemple de bonnes pratiques d'utilisation est donné par [PEREZ-RIVEROL et al., 2016]. Plus généralement, les sciences computationnelles nécessitent l'adoption de certains standards et pratiques pour assurer une bonne reproductibilité, et ceux-ci restent majoritairement à développer : [WILSON et al., 2017] donne des premières pistes. Concernant la qualité des données, de nombreux efforts sont faits pour introduire des cadres de standardisation des données : par exemple [VEIGA et al., 2017] décrit un cadre conceptuel visant à guider la résolution de problème récurrent liés à la qualité des données de biodiversité (comme par exemple évaluer des mesures jugeant de l'usage possible d'un jeu de données pour un problème donné).

L'accès aux données est également un point crucial pour la reproductibilité, et sans nous y attarder car cela impliquerait des développements sur la définition, la philosophie, le droit des données etc. qui sont des sujets de recherche en eux-même, nous donnons des perspectives sur les potentiels d'une ouverture systématique des données en recherche. En géographie, les *data paper* sont une pratique inexistante, et la règle est plutôt de garder la main jalousement sur un jeu produit, capitalisant sur le fait d'être le seul à y avoir accès. Il est évident que la qualité et quantité des connaissances produites sera nécessairement plus grande si un jeu de données est publiquement ouvert, puisqu'au moins la même chose sera obtenue, et on peut s'attendre à une prise en main par d'autres domaines, d'autres méthodes, et donc à une plus grande richesse. La fermeture induira plutôt des effets négatifs, comme par exemple du temps perdu à recoder un base vectorielle donnée uniquement sous forme de carte dans un article. L'argument du temps passé comme justification à la fermeture est absurde, puisqu'au contraire, en voyant les données comme une composante à part entière de la connaissance (voir le cadre de connaissances en 9.3), le temps passé doit impliquer plus de citations, donc plus d'utilisation, ce qui passe nécessairement par l'ouverture pour des données. De même, quelle logique, sinon la même absurde de propriété des connaissances, pousse les géographes à insérer un copyright sur l'ensemble de leurs cartes mais aussi leurs figures, jusqu'à un copyright pour un simple histogramme qui s'en serait bien passé si on avait pu

l'interroger, honnête de simplicité? Une expérience de revue induit à réellement s'inquiéter sur la valeur donnée à l'ouverture des données par les auteurs : au bout d'une dizaine d'articles, incluant des journaux affichant comme priorité et pré-requis l'ouverture totale des données et modèles, dont un seul est seulement partiellement ouvert et l'ensemble des autres implique de croire sur parole les résultats présentés (alors qu'un des buts de la revue est de contourner les biais cognitifs qu'un ou des humains ont forcément par une validation croisée qui doit se faire sur les résultats bruts et non des interprétations contenant ces biais), il est difficile de croire que des mutations profondes des pratiques ne sont pas nécessaire. Mais en suivant l'adage de Framasoft, "la route est longue mais la voie est libre", les perspectives sont nombreuses pour une évolution dont la lenteur n'est pas inéluctable. Le journal Cybergéo, pionnier des pratiques d'ouverture en sciences sociales (première revue entièrement électronique, première revue à lancer une rubrique de *model papers*), lance en 2017 une rubrique *data papers* visant à inciter le développement du partage de données et de l'ouverture en géographie. Il reste des zones grises sur lesquelles il est impossible aujourd'hui d'avoir des perspectives, notamment le droit des données. On peut citer des exemples parmi les études empiriques que nous développons : les données bibliographiques sont obtenues au prix d'une guerre de blocage par Google et un effort considérable pour la gagner; les données immobilières proviennent d'une base propriétaire achetée avec de l'argent public, et nous pouvons profiter d'un flou du contrat pour les rendre disponibles de manière agrégées avec les résultats; les données des stations essence proviennent d'une source dont la légalité ne devrait pas être creusée plus, et nous ne pouvons malheureusement pas les rendre disponibles sans prendre de risques - cet aspect n'a cependant jamais fait broncher les reviewers qui n'ont même pas mentionné le manque d'accès aux données. L'ouverture implique un engagement qui fait résolument partie de nos positionnements. C'est la même idée qui soutient la construction de l'application CybergeoNetworks⁵, qui couple les outils présentés en 2.2 avec d'autres approches complémentaires d'analyse de corpus, dans le but d'encourager la réflexivité scientifique, et de mettre cet outil ouvert à la disposition d'éditeurs indépendants, pour s'émanciper de la nouvelle main mise des géants de l'édition qui à la recherche d'un nouveau modèle pour sécuriser leur profits parient sur la vente de meta-contenu et de son analyse. Heureusement, la récente loi numérique en France a gagné le bras de fer contre leur revendication d'un droit exclusif sur la fouille de texte complets.

⁵ <http://shiny.parisgeo.cnrs.fr/CybergeoNetworks>

3.2 DONNÉES MASSIVES, CALCUL INTENSIF ET EXPLORATION DES MODÈLES

Nous nous positionnons à présent sur les questions liées à l'utilisation des données massives et du calcul intensif, ce qui induit par extension une réflexion sur les méthodes d'exploration de modèles. Il n'est pas évident que ces nouvelles possibilités soient nécessairement accompagnées de mutations épistémologiques profondes, et nous montrons au contraire que leur utilisation nécessite plus que jamais un dialogue avec la théorie. Implicitement, cette position préfigure le cadre épistémologique pour l'étude des Systèmes Complexes dont nous donnons le contexte à la section suivante 3.3 et que nous formalisons en ouverture 9.3.

3.2.1 Pour un usage raisonné des données massives et de la computation

La soi-disante *révolution des données massives* réside autant dans la disponibilité de grands jeux de données de nouveaux types variés, que dans la puissance de calcul potentielle toujours en augmentation. Même si le *tournant computationnel* ([ARTHUR, 2015]) est central pour une science consciente de la complexité et est sans doute la base des pratiques de modélisation futures en géographie comme [BANOS, 2013] souligne, nous soutenons que à la fois le *déluge de données* et les *capacités de calcul* sont dangereuses si non cadrées dans un cadre théorique et formel propre. Le premier peut biaiser les directions de recherche vers les jeux de données disponibles avec le risque de se déconnecter d'un fond théorique, tandis que le second peut occulter des résolutions analytiques préliminaires essentielles pour un usage cohérent des simulations. Nous avançons que les conditions pour la majorité des résultats dans cette thèse sont en effet ceux mis en danger par un enthousiasme inconsidéré pour les données massives, tirant la conclusion qu'un challenge majeur pour la géocomputation future est une intégration sage des nouvelles pratiques au sein du corpus existant de connaissances.

La puissance de calcul disponible semble suivre un tendance exponentielle, comme une sorte de loi de Moore. Grace à d'une part la loi de Moore effective pour le matériel, d'autre part l'amélioration des logiciels et algorithmes, conjointement avec une démocratisation de l'accès au infrastructures de simulation à grande échelle, permet à toujours plus de temps processeur d'être disponible pour le chercheur en sciences sociales (et pour le scientifique en général, mais cette mutation a déjà été opérée depuis plus longtemps dans d'autres domaines). Il y a environ une dizaine d'année, [GLEYZE, 2005] était forcé de conclure que les analyses de réseau, pour les transports publics parisiens, étaient "limitées par le calcul". Aujourd'hui la plupart des mêmes analyses seraient rapidement réglée sur un ordinateur per-

sonnel avec les logiciels et programmes appropriés : [LAGESSE, 2015] est un témoin d'un tel progrès, introduisant des nouveaux indicateurs avec une plus grande complexité de calcul, qui sont calculés sur des réseaux à grande échelle. Le même parallèle peut être fait pour les modèles Simpop : les premiers modèles Simpop au début du millénaire [SANDERS et al., 1997] étaient "calibrés" à la main, tandis que [COTTINEAU et al., 2015a] calibre le modèle Marius en multi-modélisation et [SCHMITT et al., 2014] calibre très précisément le modèle SimpopLocal, chacun sur la grille avec des milliards de simulations. Un dernier exemple, le champ de la *Space Syntax*, a témoigné d'une longue route et de progrès considérables depuis ses origines théoriques [HILLIER et HANSON, 1989] jusqu'à ses récentes applications à grande échelle [HILLIER, 2016].

Concernant les nouvelles données "massives" qui sont disponibles, il est clair que des quantités toujours plus grandes et des types toujours nouveaux sont disponibles. De nombreux exemples de champs d'application peuvent être donnés. La mobilité en est typique, puisque étudiée selon divers points de vue, comme les nouvelles données issues des systèmes de transport intelligents [O'BRIEN, CHESHIRE et BATTY, 2014], des réseaux sociaux [FRANK et al., 2014], ou des données plus exotiques comme des données de téléphonie mobile [DE NADAI et al., 2016]. Dans un autre esprit, l'ouverture de jeux de données "classiques" (comme les applications synthétiques urbaines, les initiatives gouvernementales pour les données ouvertes) devrait pouvoir toujours plus de métanalyses. De nouvelles façons de pratiquer la recherche et produire des données sont également en train d'émerger, vers des initiatives plus interactives et venant de l'utilisateur. Ainsi, [COTTINEAU, 2016] décrit une application web ayant pour but de présenter une métanalyse de la loi de Zipf sur de nombreux jeux de données, mais en particulier inclut une option de dépôt, à travers laquelle l'utilisateur peut télécharger son propre jeu de données et l'inclure dans la métanalyse. D'autres applications permettent l'exploration interactive de la littérature scientifique pour une meilleure connaissance d'un horizon scientifique complexe, comme [CHASSET et al., 2016] fait.

Comme toujours la situation n'est naturellement pas aussi idyllique qu'elle semble être au premier abord, et l'herbe verte du pré du voisin que nous pouvons être tentés d'aller brouter se transforme rapidement en un triste fumier. En effet, les objectifs et motivations sont flous et on peut facilement s'y perdre. Des illustrations parleront d'elles-mêmes. [BARTHELEMY et al., 2013] introduit un nouveau jeu de données et des méthodes relativement nouvelles pour quantifier l'évolution du réseau de rues, mais les résultats, sur lesquels les auteurs semblent s'étonner, sont qu'une transition a eu lieu à Paris à l'époque d'Haussmann. Tout historien de l'urbanisme s'interrogerait sur le but exact de l'étude, puisque à la fin un sentiment étrange de réinven-

tion de la roue flotte dans l'air. L'utilisation des ressources de calcul peut également être exagéré, et dans le cas de la modélisation multi-agent, on peut citer [AXTELL, 2016], pour lequel l'objectif de simuler le système à l'échelle 1 :1 semble être loin des motivations et justifications originelles de la modélisation agent, et pourrait même donner des arguments aux économistes *mainstream* qui dénigrent facilement les ABMS. D'autres anecdotes peuvent inquiéter : il existe en ligne des exemples étonnantes, comme une application web⁶ qui utilise des ressources de calcul financées par l'argent public pour simuler des distributions Gaussiennes afin de calculer pour un modèle de Gibrat, afin de calculer leur moyenne et variance, qui sont des paramètres d'entrée du modèle. En résumé, cela revient à vérifier le Théorème de la Limite Centrale. D'autre part, la distribution complète donnée par un modèle de Gibrat est entièrement connue théoriquement comme résolu e.g. par [GABAIX, 1999]. Sur ce point, nous devons partiellement être en désaccord avec le neuvième commandement de BANOS, qui rappelle que "les mathématiques ne sont pas le langage universel des modèles", ou plutôt souligner les dangers d'une mauvaise interprétation de ce principe⁷ : il postule que des moyens alternatifs aux mathématiques existent pour faire comprendre des processus ou des méthodes, mais précise que ceux-ci sont une porte d'entrée et ne prétend jamais qu'il est possible de se passer des mathématiques, dérive que l'exemple précédent illustre parfaitement. D'ailleurs, il est possible d'exhiber des structures mathématiques très simples, comme un simplexe en dimension quelconque, dont la visualisation "simple" est un problème ouvert. Les données fournissent aussi leur collection de dérives. Récemment, sur la liste de diffusion de géographie franco-phone *Geotamtam*, un soudain engouement autour des données issues de *Pokemon Go* a semblé répondre plus à un besoin urgent et inexpliqué d'exploiter cette source de données avant tous les autres, plutôt qu'à des considérations théoriques élaborées. Des jeux de données existant et précis, comme la population historiques des villes (pour la France la base Pumain-INED par exemple), sont loin d'être entièrement exploités et il pourrait être plus pertinent de se concentrer sur ces jeux de données classiques qui existent déjà. De même, il faut être conscient des possibles applications de résultats basée sur des malentendus : [LOUAIL et al., 2016] analyse la redistribution potentielle des transactions de carte bancaire au sein d'une ville, mais présente les résultats comme la base possible de recommandations de politiques pour une équité sociale en agissant sur la mobilité, oubliant que la forme et les fonctions urbaines sont couplés de manière complexe et que déplacer des transactions d'un endroit à un autre implique des

⁶ voir <http://shiny.parisgeo.cnrs.fr/gibratsim/>

⁷ De manière générale, les commandements de BANOS paraissent simples dans leur formulation, mais sont d'une profondeur et d'une complexité déconcertante lorsqu'on essaye d'en tirer les implications et la philosophie globale sous-jacente, et ne doivent jamais être pris à la légère.

processus bien plus complexes que des régulations directes, qui d'autant plus ne s'appliquent jamais de la façon prévue et conduisent à des résultats un peu différents. Une telle attitude, souvent observée de la part de physiciens, est très bien mise en allégorie par la figure 5 qui n'est qu'à moitié une exagération de certaines situations.

Notre principal argument est que le tournant computationnel et les pratiques de simulation seront centrales en géographie, mais peuvent également être dangereux, pour les raisons illustrées ci-dessus, i.e. que le déluge de données peut imposer les sujets de recherche et occulter la théorie, et que la computation peut éluder la construction et la résolution de modèles. Un lien plus fort est nécessaire entre les pratiques de calcul, l'informatique, les mathématiques, les statistiques et la géographie théorique. La Géographie Théorique et Quantitative est au centre de cette dynamique, puisqu'il s'agit de sa motivation initiale principale qui semble oubliée dans certains cas. Cela implique un besoin de recherche de théorie élaborées intégrées avec des pratiques de simulation conscientes. En d'autres mots, on peut répondre à des questions naïves complémentaires qui ont toutefois besoin d'être traitées une bonne fois pour toutes. Si une géographie quantitative libérée de la théorie serait possible, la réponse est naturellement non puisque cela se rapproche du piège de la fouille de données par boîte noire. Quoi qu'il soit fait par cette approche, les résultats auront un pouvoir explicatif très faible, puisqu'ils pourront mettre en valeur des relations mais pas reconstruire des processus. D'autre part, la possibilité d'une géographie quantitative purement basée sur le calcul est une vision dangereuse : même le gain de trois ordres de grandeur dans la puissance de calcul disponible ne résout pas le sort de la dimension. Prenons l'exemple des résultats de non-stationnarité obtenus en 4.1. L'utilisation de données relativement massives, de par les algorithmes spécialement conçus pour être capable de faire les traitements, est une condition nécessaire au résultat obtenus, mais à la fois l'échelle est les objets (c'est à dire les indicateurs calculés) sont co-déterminés par les constructions théoriques et les autres études empiriques. En effet l'absence de théorie impliquerait de ne pas connaître les objets, mesures et propriétés à étudier (e.g. le caractère multi-scalaire ou dynamique des processus), et sans résolutions analytiques, il serait souvent difficile de tirer des conclusions à partir des analyses empiriques seules concernant l'ergodicité par exemple. Rien n'est vraiment nouveau ici mais cette position doit être affirmée et tenue, précisément car notre travail se base sur ce type d'outils, essayant d'avancer sur une arête fine et fragile, avec d'un côté le vide du charlatanisme théorique infondé et de l'autre l'abîme de l'overdose technocratique dans des quantités de données folles. Plus que jamais on a besoin de théories simples mais fondées et puissantes à-la-Occam [BATTY, 2016], pour permettre une intégration saine des nouvelles techniques au sein des connaissances existantes.

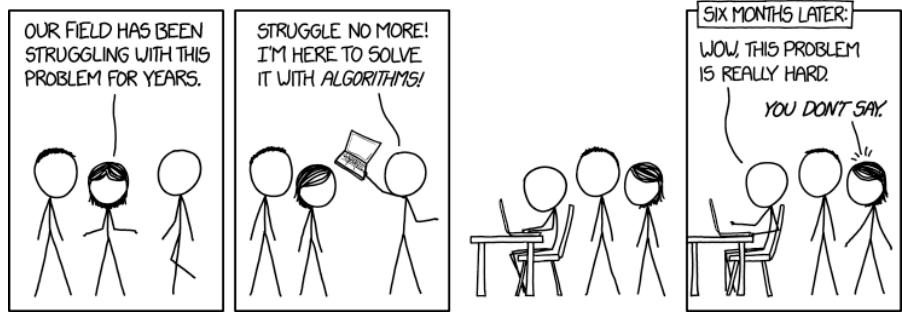


FIGURE 5 : De l'usage naïf de la fouille de données et du calcul intensif.
Source : xkcd

3.2.2 Contrôle statistique pour les conditions initiales par génération de données synthétiques

Contexte

Lors de l'évaluation de modèle basés sur les données, ou même de modèle plus simples partiellement basés sur les données impliquant une paramétrisation simplifiée, une issue inévitable est le manque de contrôle sur les “paramètres implicites du systèmes” (ce qui n'est pas une notion stricte mais doit être vu dans notre sens comme les paramètres régissant la dynamique). En effet, une statistique issue d'executions du modèle sur un nombre suffisant d'executions peut toutefois rester biaisée, au sens où il est impossible de savoir si les résultats sont dus aux processus que le modèle cherche à traduire ou à une structure présente dans les données initiale. La question méthodologique fondamentale qui nous intéressera pour la suite est d'être capable d'isoler les effets propres aux processus du modèles de ceux liés à la géographie.

RATIONNELLE Bien que les modèles de simulation des systèmes géographiques en général et les modèles basés-agent en particulier représentent une opportunité considérable d'explorer les comportements socio-spatiaux et de tester une variété de scenarios pour les politiques publiques, la validité des modèles génératifs est incertaine tant que la robustesse des résultats n'a pas été établie. Les analyses de sensibilité incluent généralement l'analyse des effets de la stochasticité sur la variabilité des résultats, ainsi que les effets de variations locales des paramètres. Cependant, les conditions spatiales initiales sont généralement prise pour données dans les modèles géographiques, laissant ainsi totalement inexploré l'effet des motifs spatiaux sur les interactions des agents et sur leur interaction avec l'environnement. Dans cette partie, nous présentons une méthode pour établir l'effet des conditions spatiales initiales sur les modèles de simulation, utilisant un générateur systématique contrôlé par des meta-

paramètres pour créer des grilles de densité utilisées dans les modèles de simulation spatiaux. Nous montrons, avec l'exemple d'un modèle agent très classique (le modèle Sugarscape d'extraction de ressources) que l'effet de l'espace dans les simulations est significatif, et parfois plus grand que l'effet des paramètres eux-mêmes. Nous y arrivons en utilisant le calcul haute performance en un workflow très simple et open source. Les bénéfices de notre approche sont variés mais incluent par exemple la connaissance du comportement du modèle dans un contexte plus large, la possibilité de contrôle statistique pour régresser les sorties du modèle, ou une exploration plus fine des dérivées du modèle que par rapport à une approche directe.

FORMALISATION Commençons par donner une formulation abstraite de l'idée, d'un point de vue du couplage de modèle. Le générateur est considéré comme un modèle amont, couplé simplement (les sorties devenant les entrées) avec le modèle aval étudié. Si M_u est le modèle amont, M_d le modèle aval et α les meta-paramètres, on a la composition de la dérivée le long des meta-paramètres

$$\partial_\alpha [M_u \circ M_d] = (\partial_\alpha M_u \circ M_d) \cdot \partial_\alpha M_d$$

Cela implique que la sensibilité du modèle aval aux meta-paramètres peut être déterminée en étudiant le couplage séquentiel et le modèle amont. Nous gagnons de la connaissance thématique, dans la sensibilité à un meta-paramètre implicite, mais il y a aussi un gain computationnel : la génération de différentielles contrôlées dans l'espace initial (c'est à dire tester directement la comparaison entre deux grilles proches) serait compliquer à atteindre directement. La question de la stochasticité dans de tels modèles couplés simplement ne pose pas de problème supplémentaire puisque $E[X] = E[E[X|Y]]$. Cela multiplie naturellement le nombre de répétitions pour converger bien évidemment. Nous resterons dans l'application pratique ici à une étude de l'espace faisable de sortie et non à une étude différentielle, cette considération théorique n'influe pas à cet ordre, mais doit être gardée à l'esprit pour d'éventuelles applications plus fines.

ROLE DE LA DÉPENDANCE AU CHEMIN SPATIO-TEMPORELLE La dépendance au chemin spatio-temporelle est une des raisons principales rendant notre approche pertinente. En effet, un aspect crucial de la plupart des systèmes complexes spatio-temporels est leur non-ergodicité [PUMAIN, 2012b] (la propriété que les échantillons cross-sectionnels dans l'espace ne sont pas équivalents aux échantillons dans le temps pour calculer des statistiques comme la moyenne), qui témoigne généralement de forte dépendances au chemin spatio-temporelles dans les trajectoires. De manière similaire à ce que GELL-MANN appelle *frozen accidents* dans tout système complexe [GELL-MANN, 1995], une configuration donnée contient des indices sur les bifurcations

passées, qui peuvent avoir eu des effets considérables sur l'état du système. Les effets temporels et cumulatifs ont été considérés dans de nombreux sous-champs géographiques et à différentes échelles géographiques, par exemple les systèmes régionaux [WILSON, 1981] ou l'échelle intra-urbaine [ALLEN et SANGLIER, 1979]. L'impact de la configuration spatiale sur les dynamiques du modèle et les bifurcations spatiales a été moins étudié.

L'exemple des réseaux de transport est une bonne illustration, car leur forme spatiale et leur hiérarchie est fortement influencée par les décisions d'investissement du passé, les choix techniques, ou des décisions politiques qui ne sont parfois pas rationnelles [ZEMBRI, 2010]. Certains indicateurs agrégés ne prendront pas en compte les positions et trajectoires de chaque agent (comme les inégalités totales dans le modèle Sugarscape) mais d'autres, comme dans le cas des motifs d'accèsibilité spatiale dans un système de villes, capture entièrement la dépendance au chemin et peuvent ainsi être fortement dépendants à la configuration spatiale initiale. Il n'est pas clair par exemple ce qui a causé la transition de la capitale française de Lyon à Paris dans le bas Moyen-Age, certaines hypothèses étant la reconfiguration des motifs commerciaux du Sud au Nord de l'Europe et donc une centralité accrue pour Paris due à sa position spatiale, tout en gardant à l'esprit que les centralité géographique et politique ne sont pas équivalentes et entretiennent une relation complexe [GUENÉE, 1968]. La bifurcation induite par des facteurs socio-économiques et politiques a pris une signification profonde avec des répercussions mondiales encore aujourd'hui quand elle a été concrétisée par la configuration spatiale.

TRAVAUX EXISTANTS L'effet de la configuration spatiale sur les attributs agrégés à la zone des comportements humains a été largement discuté en géostatistiques, approximativement depuis l'introduction du *Modifiable Areal Unit Problem* (MAUP) [OPENSHAW, 1984]. Plus récemment, [KWAN, 2012] plaide pour un examen plus attentif de ce qui serait un *Uncertain Geographic Context Problem* (UGCoP), qui est la configuration spatiale des unités géographiques même si la taille et la délimitation des zones est la même. Au contraire, le faible nombre de considérations similaires dans la littérature traitant des modèles de simulation géographiques remet en question la généralisation de leur résultats, comme cela a été montré par exemple dans le cas des modèles LUTI [THOMAS et al., 2017], ou des processus de diffusion étudiés par modèles basé-agents [LE TEXIER et CARUSO, 2017].

Méthodes

Nous détaillons à présent la méthode développée pour analyser la sensibilité des modèles de simulation aux conditions spatiales initiales. S'ajoutant au protocole usuel, qui consiste à simuler un mo-

dèle μ pour différentes valeurs de ses paramètres et faire le lien entre ces variations aux variations des résultats de simulation, nous introduisons ici un générateur spatial, qui est lui-même déterminé par des paramètres et produit des ensembles de configurations spatiales initiales. Les configurations spatiales initiales sont catégorisées pour représenter des types d'espace typiques (par exemple des grilles de densité monocentriques ou polycentriques), et la sensibilité du modèle est à présent testée sur les paramètres de μ mais aussi sur les paramètres spatiaux ou les types spatiaux. Cela permet à l'analyse de sensibilité de fournir des conclusions qualitatives au regard de l'influence de la distribution spatiale sur les sorties des modèles de simulation, en parallèle des variation classiques des paramètres.

GÉNÉRATEUR SPATIAL Le générateur spatial applique un modèle de morphogenèse urbaine développé et exploré en 6.2. Pour le présenter rapidement, les grilles sont générées par un processus itératif qui ajoute une quantité de population N à chaque pas de temps, l'allouant selon un attachement préférentiel caractérisé par sa force d'attraction α . The premier processus est ensuite lissé n fois par un processus de diffusion de force β . Les grilles sont donc générées aléatoirement par la combinaison des valeurs de ces quatre meta-paramètres α , β , n and N . Pour faciliter l'exploration, seule la distribution de densité est autorisée à varier plutôt que la taille de la grille, qui est fixée à un environnement carré 50x50 de population 100,000 unités.

COMPARER LES DIAGRAMMES DE PHASE Afin de tester l'influence des conditions spatiales initiales, nous avons besoin d'une méthode systématique pour comparer des diagrammes de phase. En effet, nous avons autant de diagramme de phase que de grilles spatiales, ce qui rend une comparaison visuelle qualitative non réaliste. Une solution est d'utiliser des procédures quantitatives systématiques. De nombreuses méthodes pourraient potentiellement être utilisées : par exemple, des indicateurs anisotropes comme la donnée de clusters et leur position dans le diagramme de phase, peuvent permettre de révéler des *meta-transitions de phase* (transition de phase dans l'espace des meta-paramètres. L'utilisation de métriques comparant des distributions spatiales, comme la *Earth Movers Distance* qui est utilisée en viion par ordinateur pour comparer des distributions de probabilité [RUBNER, TOMASI et GUIBAS, 2000], ou la comparaison de matrices de transition agrégées de la dynamique associée au potentiel décrit par chaque distribution, est également possible. Les méthodes de comparaison de cartes, répandues en sciences environnementales, fournissent de nombreux outils pour comparer des champs en deux dimensions [VISSER et DE NIJS, 2006]. Pour comparer un champ spatial évoluant dans le temps, des méthodes élaborées comme les Fonctions Orthogonales Empiriques qui isolent les variations temporelles

des variations spatiales, seraient applicables dans notre cas en prenant le temps comme une dimension de paramètre, mais celles-ci ont été montrées ayant une performance similaire à la comparaison visuelle directe lorsqu'on prend la moyenne sur un ensemble de contributions crowdsourcées [KOCH et STISEN, 2017]. Pour rester simple et car de telles considérations méthodologiques sont auxiliaires pour le propos principal de cette partie, nous proposons une mesure intuitive correspondant à la part de la variabilité inter-diagrammes relativement à leur variabilité interne. Plus formellement, cette distance est donnée par

$$d_r(\alpha_1, \alpha_2) = 2 \cdot \frac{d(f_{\vec{\alpha}_1}, f_{\vec{\alpha}_2})^2}{\text{Var}[f_{\vec{\alpha}_1}] + \text{Var}[f_{\vec{\alpha}_2}]} \quad (1)$$

où $\alpha \mapsto [\vec{x} \mapsto f_{\vec{\alpha}}(\vec{x})]$ est l'opérateur donnant les diagrammes de phase avec \vec{x} paramètres et $\vec{\alpha}$ meta-paramètres, et d une distance entre distributions de probabilité qui peut être prise par exemple comme la distance L2 basique ou la *Earth Movers Distance*. Pour chaque valeur $\vec{\alpha}_i$, le diagramme de phase est vue comme un champ spatial aléatoire, ce qui facilite la définition des variances et de la distance.

Résultats

Sugarscape est un modèle d'extraction de ressources qui simule la distribution inégale des richesses dans une population hétérogène [EPSTEIN et AXTELL, 1996]. Des agents ayant différentes portées de vision et différents métabolismes collectent une ressource qui se régénère automatiquement et disponible de manière hétérogène dans le paysage initial. Ceux-ci s'établissent et collectent la ressource, ce qui mène certains d'entre eux à survivre et d'autres à périr. Les paramètres principaux du modèle sont le nombre d'agents, leur ressources minimale et maximale. Nous nous intéressons en prime à tester l'impact de la distribution spatiale, en utilisant le générateur spatial. La sortie du modèle est mesurée comme le diagramme de phase d'un index d'inégalité pour la distribution de la ressource (index de Gini). Nous étendons l'implémentation ayant initialement une distribution de richesse des agents, donnée par [LI et WILENSKY, 2009].

Pour l'exploration, 2,500,000 simulations (1000 points de paramètres \times 50 grilles de densité \times 50 réplications) nous permettent de montrer que le modèle est bien plus sensible à l'espace qu'à ses autres paramètres, à la fois quantitativement et qualitativement : l'amplitude des variations entre les grilles de densité est plus grande que l'amplitude dans chaque diagramme de phase, et le comportement de ces diagrammes de phase est qualitativement différents dans diverses régions de l'espace morphologique. Plus précisément, nous explorons une grille d'un espace de paramètre basique du modèle, dont les trois dimensions sont la population des agents $P \in [10; 510]$, la ressource

minimale initiale par agent $s_- \in [10; 100]$ et la ressource initiale maximale par agent $s_+ \in [110; 200]$. Chaque paramètre est discréteisé en 10 valeurs, donnant 1000 points de paramètres. Nous procédons à 50 répétitions pour chaque configuration, ce qui donne des propriétés de convergence raisonnables. La distribution spatiale initiale varie parmi 50 grilles initiales, générée en échantillonnant les meta-paramètres du générateur dans un Hypercube Latin. Nous démontrons ainsi la flexibilité de notre cadre, par le couplage séquentiel direct du générateur avec le modèle. Nous mesurons la distance de l'ensemble des diagrammes de phase à 3 dimensions à un diagramme de phase de référence calculé sur l'initialisation du modèle par défaut (voir Fig. 6 pour sa position morphologique au regard des grilles générées), en utilisant l'équation 1 avec la distance L2 pour assurer une interprétabilité directe. En effet, cela donne dans ce cas la distance au carré moyen entre chaque points en correspondance des diagrammes, relative à la moyenne des variances de chaque. Pour cela, des valeurs plus grandes que 1 signifient que la variabilité inter-diagramme est plus importante que la variabilité intra-diagramme.

Nous obtenons une sensibilité très forte aux conditions initiales, puisque la distribution de la distance relative à la référence s'étend sur l'ensemble des grilles de 0.09 à 2.98, avec un médiane de 1.52 et une moyenne de 1.30. Cela signifie qu'en moyenne, le modèle est plus sensible aux meta-paramètres qu'aux paramètres, et que la variation relative peut atteindre jusqu'à un facteur 3. Nous montrons en Fig. 6 leur distribution dans un espace morphologique. L'espace morphologique réduit est obtenu en calculant 4 indicateurs bruts de forme urbaine, qui sont l'index de Moran, la distance moyenne, le niveau de hiérarchie et l'entropie (voir [LE NÉCHET, 2015] ainsi que la section 6.2 pour une définition précise et une mise en contexte), et en réduisant la dimension avec une analyse par composantes principales pour laquelle nous gardons les deux premières composantes (92% de variance cumulée). La première mesure un "niveau d'étalement" et d'éclatement, tandis que la seconde mesure l'agrégation.⁸ Nous trouvons que les grilles produisant les déviations les plus grandes sont celles avec un faible niveau d'étalement et une forte agrégation. Cela est confirmé par le comportement comme fonction des meta-paramètres, puisque des fortes valeurs de α donnent aussi une forte distance. En terme de processus du modèle, cela montre que les mécanismes de congestion induisent rapidement de plus haut niveau d'inégalités.

Nous contrôlons à présent la sensibilité en terme de comportement qualitatif des diagrammes de phase. Nous montrons en Fig. 7 les diagrammes pour deux morphologies très opposées en terme d'étalement, mais en contrôlant l'agrégation par la même valeur de PC2.

⁸ nous avons $PC1 = 0.76 \cdot \text{distance} + 0.60 \cdot \text{entropy} + 0.03 \cdot \text{moran} + 0.24 \cdot \text{slope}$ et $PC2 = -0.26 \cdot \text{distance} + 0.18 \cdot \text{entropy} + 0.91 \cdot \text{moran} + 0.26 \cdot \text{slope}$.

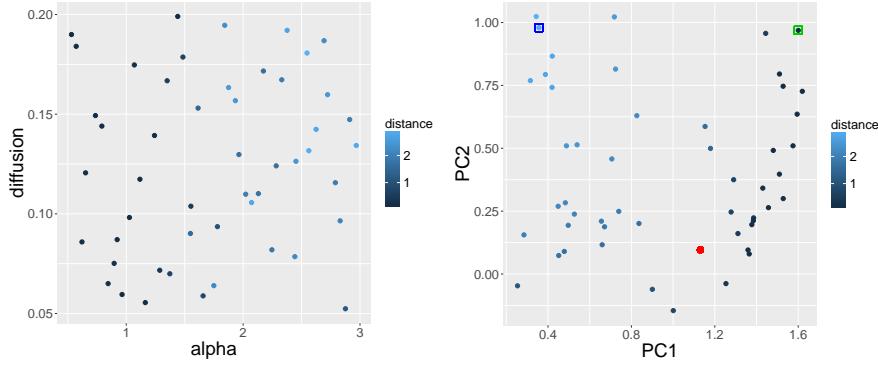


FIGURE 6 : Distance relative des diagrammes de phase à la référence pour l’ensemble des grilles. (Gauche) Distance relative comme fonction des meta-paramètres α (force de l’attachement préférentiel) et la diffusion (β , force du processus de diffusion). (Droite) Distance relative comme fonction des deux composantes principales de l’espace morphologique (voir texte). Le point rouge correspond à la configuration spatiale de référence. Les cadres verts et bleu donnent respectivement le premier et le second diagrammes particuliers montrés à la Fig. 7.

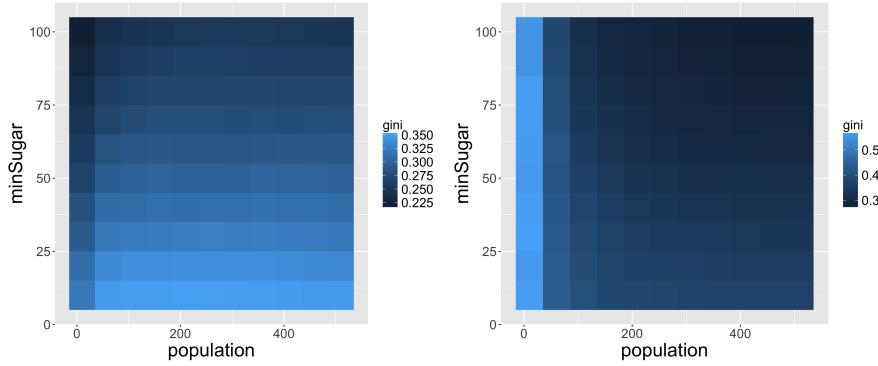


FIGURE 7 : Exemples de diagrammes de phase. Nous montrons deux diagrammes bi-dimensionnels sur (P, s_-) , obtenus à $s_+ = 110$ fixé. (Gauche) Cadre vert, obtenu avec $\alpha = 0.79$, $n = 2$, $\beta = 0.14$, $N = 157$; (Droite) Cadre bleu, obtenu avec $\alpha = 2.56$, $n = 3$, $\beta = 0.13$, $N = 128$.

Ceux-ci correspondent au cadres vert et bleu en Fig. 6. Les comportements sont relativement stables pour s_+ variant, ce qui signifie que les agents les plus pauvres ont un rôle déterminant dans les trajectoires. Les deux exemples ont non seulement une inégalité de base très distance (le plafond du premier 0.35 est environ le plancher du second 0.3), mais leur comportement qualitatif est également radicalement opposé : la configuration étalée donne des inégalités qui décroissent quand la population décroît et qui décroissent quand la richesse minimale augmente, tandis que la concentrée donne des inégalités augmentant fortement quand la population décroît et aussi décroissantes avec la richesse minimale mais significativement seulement pour des grandes valeurs de population. Le processus est ainsi complètement inversé, ce qui aurait un impact déterminant si l'on essayait de schématiser des politiques à partir du modèle. Cet exemple confirme ainsi l'importance de la sensibilité des modèles de simulation aux conditions spatiales initiales.

3.2.3 *Lien entre modélisation et Science Ouverte*

Enfin, il est important de souligner brièvement les liens entre pratiques de modélisation et science ouverte, comme le lien entre reproductibilité et science ouverte souligné à la fin de 3.1. En fait, la Science Ouverte est composée d'un ensemble de pratiques sur différents points, d'où sa ventilation logique dans nos positionnements. Pour illustrer les enjeux, nous proposons de décrire l'exemple des workflows d'exploration de modèle comme une méthode de meta-analyse de sensibilité, c'est à dire un aspect de la méthodologie appliquée ci-dessus. Les idées de multi-modélisation et d'exploration intensive de modèle sont tout sauf nouvelles puisque OPENSHAW défendait déjà le "model-crunching" dans [OPENSHAW, 1983], mais leur utilisation effective commence seulement à émerger grâce à l'apparition de nouvelles méthodes et outils en même temps qu'une explosion des capacités de calcul : [COTTINEAU, REY et REUILLOU, 2016] plaide pour une approche renouvelée de la multi-modélisation. Le couplage de modèles comme nous faisons répond à des questions similaires. Dans cette lignée de recherche, la plateforme d'exploration de modèle OpenMole [REUILLOU, LECLAIRE et REY-COYREHOURCQ, 2013] permet d'embarquer n'importe quel modèle comme une boîte noire, d'écrire des workflow d'exploration modulables qui utilisent des méthodologies d'exploration avancées comme des algorithmes génétiques, et de distribuer de manière transparente les calculs sur des infrastructures de calcul à grande échelle comme des clusters ou grilles de calcul. Dans le cas précédent, l'outil du workflow est un outil puissant pour intégrer à la fois l'analyse de sensibilité et la meta-analyse de sensibilité, et permet de coupler n'importe quel générateur avec n'importe quel modèle de façon très directe tant que le modèle peut prendre

sa configuration spatiale comme entrée ou dans un fichier d'entrée. D'autre part, une idée des workflow est de favoriser des constructions ouvertes et collaboratives, puisque le “marketplace” d'OpenMole, directement intégré au logiciel, permet de bénéficier directement des exemples qui auront été partagés sur le dépôt collaboratif. Cela ressemble aux plateformes de partage de modèles, qui sont nombreuses pour les modèles agents par exemple, mais dans un esprit encore plus modulaire et participatif. Ainsi, certains choix épistémologiques et méthodologiques au regard de la modélisation impliquent directement un positionnement au regard de la science ouverte : la multimodélisation et les familles de modèles, qui vont de pair avec le couplage de modèle hétérogènes et multi-échelles, ne peuvent guère être viables sans des pratiques d'ouverture, de partage et de construction collaborative des modèle, comme le rappelle [BANOS, 2013].

★ ★

★

3.3 POSITIONNEMENT EPISTÉMOLOGIQUE

3.3.1 Approche cognitive et Perspectivisme

Notre positionnement épistémologique se fonde sur une approche cognitive de la science, donnée par GIERE dans [GIERE, 2010b]. L'approche se concentre sur le rôle des agents cognitifs comme porteurs et producteurs de la connaissance. Elle a été montrée opérationnelle par [GIERE, 2010a] qui étudie un modèle basé-agent de la science. Ces idées convergent avec le jeu Nobel de CHAVALARIAS [CHAVALARIAS, 2016] qui teste de manière stylisée l'équilibre entre exploration et falsification dans l'entreprise scientifique collective. Ce positionnement épistémologique a été présenté par GIERE comme *perspectivisme scientifique* [GIERE, 2010c], dont la caractéristique principale est de considérer toute entreprise scientifique comme une *perspective* dans laquelle des *agents* utilisent des *media* (modèles) pour représenter quelque chose dans un certain but. Pour concrétiser, nous pouvons le positionner sur la "check-list" du constructivisme de HACKING [HACKING, 1999], un outil pratique pour positionner une position épistémologique dans un espace simplifié à trois dimensions dans lequel les dimensions sont différents aspects sur lesquels les approches réalistes et constructivistes généralement divergent : d'abord la contingence (dépendance au chemin du processus de construction de connaissances) est nécessaire l'approche perspectiviste qui est pluraliste, deuxièmement le "degré de constructivisme" est assez haut car les agents produisent la connaissance, et enfin la stabilité des théories dépend des interactions complexes entre les agents et leur perspectives. Cela a pour ces raisons été présenté comme un chemin intermédiaire et alternatif entre le réalisme absolu et le constructivisme sceptique [BROWN, 2009]. la notion de *perspective* jouera un rôle fondamental dans le cadre développé en 9.3.

Cette approche mettant l'emphase sur l'auto-organisation, nous la voyons totalement compatible avec une vision anarchiste de la science comme défendue par FEYERABEND [FEYERABEND, 1993]. Celui-ci émet des doutes sur l'intérêt de l'anarchisme politique mais introduit l'*anarchisme scientifique*, qu'il ne faut pas comprendre comme un refus total de toute méthode "objective", mais d'une autorité et légitimité artificielle que certaines méthodes ou courants scientifique pourraient vouloir prendre. Il démontre par une analyse précise des travaux de Galilée que la plupart de ces résultats étaient basés sur des croyances et que la plupart n'étaient pas accessibles avec les outils et méthodes de l'époque, et postule qu'il devrait en être de même pour certains travaux contemporains. Il n'y a donc pas de *perspective* objectivement plus légitimes que d'autres dans la mesure de leurs validation par des faits et des pairs - et même dans ces cas la légitimité doit pouvoir être discutée, car la remise en question est un fondement de

la connaissance. Cela correspond exactement à la pluralité des perspectives que nous défendons. L'auto-organisation et l'émergence des connaissances nécessite un certain anarchisme pour échapper aux pré-conceptions cadrant par le haut. En effet, les positions anarchistes ont trouvé un écho très cohérent dans les différents courants de la complexité, de la cybernétique à l'auto-organisation au cours du 20ème siècle [DUDA, 2013]. Notre cadre de connaissance développé en 9.3 illustre cette émergence de la connaissance. De plus, notre volonté de réflexivité et de donner à notre travail des pistes de lecture diverse au delà de la linéarité (voir F), illustre l'application de ces principes. Les recommandations méthodologiques et positionnements donnés précédemment dans ce chapitre pourraient sonner comme totalitaires s'ils étaient assénés de manière sèche sans contexte, mais ceux-ci sont en fait tout le contraire puisqu'ils découlent d'un dynamique récente de science ouverte qui a bien émergé par le bas, partiellement conséquence de l'ouverture et de la pluralité.

3.3.2 *De la Vie à la Culture*

Le parallèle entre les systèmes sociaux et les systèmes biologiques est souvent fait, parfois de manière plus qu'imagée comme par exemple pour la théorie du *Scaling* de WEST qui applique des équations de croissance similaires à partir des lois d'échelle, avec des conclusions inverses tout de même concernant la relation entre taille et rythme de vie [BETTENCOURT et al., 2007]. Les relations d'échelle ne tiennent plus lorsqu'on essaye de les appliquer à une fourmi seule, et il faut alors l'appliquer à la fourmilière entière qui est alors l'organisme en question. En ajoutant la propriété de cognition, on confirme qu'il s'agit du niveau pertinent, puisque celle-ci possède des propriétés cognitives avancées, comme la résolution de problèmes d'optimisation spatiaux, ou la réponse rapide à une perturbation extérieure. Les organisations sociales humaines, les villes, peuvent-elles être vues comme des organismes ? BANOS file dans [BANOS, 2013] la métaphore de la *fourmilière urbaine* mais rappelle que le parallèle s'arrête assez vite. Nous allons voir cependant dans quelle mesure certains concepts de l'épistémologie de la biologie peuvent être utiles pour comprendre les systèmes sociaux que nous nous proposons d'étudier. Nous nous basons sur la contribution fondamentale de MONOD dans [MONOD, 1970], qui tente de développer les principes épistémologiques cruciaux pour l'étude du vivant. Ainsi, les organismes vivants répondent à trois propriétés essentielles qui permettent de les différencier d'autres systèmes : (i) la téléonomie, c'est à dire qu'il s'agit "d'objets doués d'un projet", projet qui se reflète dans leur structure et dans celles des artefacts qu'ils produisent⁹; (ii) l'importance des processus mor-

⁹ à ne pas confondre avec la téléologie, propres aux animismes, qui consiste à prêter un projet ou un sens à l'univers

phogénétiques dans leur constitution (voir 6.1); (iii) la propriété de reproduction invariante de l'information définissant leur structure. MONOD esquisse de plus en conclusion des pistes pour une théorie de l'évolution culturelle. La téléconomie est essentielle dans les structures sociales, puisque toute organisation essaye de satisfaire un ensemble d'objectifs, même si en général elle n'y parviendra pas et que ceux-ci co-évolueront avec l'organisation. Un aspect divergent est cette notion de multi-objectif qui est typique des systèmes complexes socio-techniques. Ensuite, nous postulons que la notion de morphogenèse est un outil essentiel pour comprendre ces systèmes, avec une définition très proche de celle utilisée en biologie. Un travail approfondi pour donner cette définition est fait en 6.1, que nous résumons en l'existence de processus relativement autonomes guidant la croissance du système et impliquant des relations causales circulaires entre forme et fonction qui témoignent d'une architecture émergente. Pour des systèmes sociaux, isoler le système est plus difficile et la notion de frontière sera moins stricte que pour un système biologique, mais on retrouvera bien ce lien entre forme et fonction, comme par exemple la structure d'une organisation ayant un impact sur ses fonctionnalités. Enfin, la reproduction de l'information est au cœur de l'évolution culturelle, par la transmission de la culture et la *mémétique*, la différence étant que le rapport d'échelle de temps entre la fréquence de transmission et les processus de croisement et de mutation ou d'autres processus non mémétiques de production culturelle est très faible, alors qu'elle est de plusieurs ordres de magnitude en biologie. [GABORA et STEEL, 2017] propose un modèle de réseau autocatalytique pour la cognition, qui expliquerait l'apparition de l'évolution culturelle par des processus analogues à ceux s'étant produit à l'apparition de la vie, c'est à dire une transition permettant au molécules de s'auto-entretenir et s'auto-reproduire, les représentations mentales faisant office de molécules. Cet exemple montre bien que le parallèle n'est pas toujours absurde. Mais si les processus à l'origine sont analogues, la nature de l'évolution est bien différente par la suite, comme le montre [LEEUW, LANE et READ, 2009], les critères darwiniens d'évolution n'étant pas suffisant pour expliquer l'évolution de nos sociétés organisées. Il s'agit d'un degré de complexité supérieur et le rôle des flux d'information est crucial (voir le rôle de la complexité informationnelle dans la sous-section suivante). Enfin, l'un des points sur lequel il s'agit d'être attentif, est la plus grande difficulté de définir les niveaux d'émergence pour les systèmes sociaux : [ROTH, 2009] souligne le risque de tomber dans des cul-de-sac ontologiques car les niveaux ont été mal définis, et qu'il faut d'une manière générale penser au-delà de la seule dichotomie micro-macro qui est utilisée pour caricaturer les notions d'émergence faible, mais que les ontologies doivent souvent être multi-niveaux et impliquant de multiples niveaux intermédiaires.

3.3.3 *Nature de la Complexité et Production de Connaissances*

Un aspect de la production de connaissance sur des Systèmes Complexes, auquel nous nous heurtons plusieurs fois ici (voir chapitre 9), et qui semble être récurrent voire inévitable, est une certaine réflexivité. Nous entendons par là à la fois une réflexivité pratique, c'est à dire la nécessité d'élever le niveau d'abstraction, comme le besoin de reconstruire de manière endogène les disciplines dans lesquelles une réflexion cherche à se positionner comme proposé en 2.2, ou de réfléchir à la nature épistémologique de la modélisation lors de l'élaboration d'un modèle comme en 9.2, mais également une réflexivité théorique en le sens que les appareils théoriques ou les concepts produits peuvent s'appliquer de manière récursive à eux-mêmes. Cette constatation pratique fait écho à des débats épistémologiques anciens questionnant la possibilité d'une connaissance objective de l'univers qui serait indépendante de notre structure cognitive, ou bien la nécessité d'une "rationalité évolutive" impliquant que notre système cognitif, produit de l'évolution, reflète les processus complexes ayant conduit à son émergence, et que toute structure de connaissance sera par conséquent réflexive¹⁰. Nous ne prétendons pas ici apporter une réponse à une question aussi vaste et vague telle quelle, mais proposons un lien potentiel entre cette réflexivité et la nature de la complexité.

COMPLEXITÉ ET COMPLEXITÉS Ce qui est entendu par complexité d'un système mène souvent à des malentendus car celle-ci peut être qualifiée selon différentes dimensions et visions. Nous distinguons d'une part la complexité au sens d'émergence faible et d'autonomie entre les différents niveaux d'un système, et sur laquelle différentes positions peuvent être développées comme dans [DEFFUANT et al., 2015]. Nous ne rentrerons pas dans une granularité plus fine, la vision de la complexité sociale donnant encore plus de fil à retordre au démon de Laplace, peut être par exemple comprise par une émergence plus forte, la nature des systèmes ne jouant pas de rôle dans notre reflexion. D'autre part, nous distinguons deux autres "types" de complexité, la complexité computationnelle et la complexité informationnelle, qui peuvent être vues comme des mesures de complexité, mais qui ne sont pas directement équivalentes à l'émergence, puisqu'il n'existe pas de lien systématique entre les trois. On peut par exemple imaginer utiliser un modèle de simulation, pour lequel les interactions entre agents élémentaires se traduisent par un message codé au niveau supérieur : il est alors possible en exploitant les degré de liberté de minimiser la quantité d'information contenue dans le message (ce qui serait en pratique inutile car il y a

¹⁰ Nous remercions D. Pumain d'avoir pointé cette vue alternative du problème que nous allons développer par la suite

des moyens plus simples de simuler un bruit blanc). Les différentes langues demandent des efforts cognitifs différents et compressent différemment l'information, ayant différents niveau de complexité mesurables [FEBRES, JAFFÉ et GERSHENSON, 2013]. De même, des artefacts architecturaux sont le résultat d'un processus d'évolution naturelle puis culturelle et peuvent témoigner plus ou moins de cette trajectoire. Ainsi, les liens entre ces trois types de complexité ne sont pas systématiques, et dépendent du type de système. Des liens épistémologiques peuvent néanmoins être introduits. Nous traitons ceux entre émergence et les deux autres complexités, étant donné que le lien entre complexité computationnelle et complexité informationnelle est assez bien compris et relève de problématiques de compression de l'information et de traitement du signal, ou encore de cryptographie.

COMPLEXITÉ COMPUTATIONNELLE ET ÉMERGENCE Le "paradoxe" du chat de Schrödinger n'en est un que si l'on prend une vision réductionniste, c'est à dire si l'on suppose que la superposition d'états peut se propager à travers les niveaux successifs et qu'il n'y aurait pas émergence, c'est à dire constitution d'un niveau supérieur autonome. Cette vision intuitive a récemment été démontrée rigoureusement par [BOLOTIN, 2014] qui prouve que l'acceptation de $P \neq NP$ implique une séparation qualitative entre le niveau quantique microscopique et le niveau d'observation macroscopique. En d'autres termes, la complexité computationnelle est suffisante pour avoir émergence. A priori, cette séparation effective des échelles n'implique pas que le niveau inférieur ne joue pas un rôle crucial, puisque [VATTAY, SALAHUB et CSABAI, 2015] prouve que les propriétés de criticalité quantiques sont typiques des molécules du vivant, sans qu'il n'y ait a priori de spécificité pour la vie dans cette détermination complexe par les échelles inférieures : [VERLINDE, 2016] a introduit une nouvelle approche liant théories quantiques et relativité générale dans laquelle il est montré que la gravité est un phénomène émergent et que la dépendance au chemin dans la déformation de l'espace de base introduit un terme supplémentaire au niveau macroscopique, qui permet d'expliquer les déviations attribuées jusqu'alors à la "matière noire". Dans le sens inverse, le lien entre complexité computationnelle et émergence est mis en valeur par les questions liées à la nature de la computation [MOORE et MERTENS, 2011]. Des automates cellulaires, qui sont par ailleurs cruciaux pour la compréhension de divers systèmes complexes, ont été montrés Turing-complets (comme le Jeu de la Vie). Des organismes sans système nerveux central sont capables de résoudre des problèmes difficiles [REID et al., 2016]. Ce lien fondamental avait été envisagé par TURING, puisqu'au delà de ses contributions fondamentales à l'informatique moderne, il s'était intéressé à la morphogenèse et a tenté de produire des modèles chimiques d'explication de celle-ci [TURING, 1952] (qui étaient très loin

d'effectivement de l'expliquer - elle n'est toujours pas bien comprise aujourd'hui, voir [6.1](#) - mais dont les contributions conceptuelles ont été fondamentales, notamment pour la notion de réaction-diffusion).

COMPLEXITÉ INFORMATIONNELLE ET ÉMERGENCE La complexité informationnelle, ou la quantité d'information contenue dans un système et la manière dont celle-ci est stockée, entretient également des liens fondamentaux avec l'émergence. L'information est équivalente à l'entropie d'un système et donc à son degré d'organisation - c'est ce qui a permis de résoudre le paradoxe apparent du Démon de Maxwell qui serait capable de diminuer l'entropie d'un système isolé et donc contredire la deuxième loi de la thermodynamique : celui-ci utilise en fait l'information sur les positions et vitesses des molécules du système, et son action compense la perte d'entropie par sa captation d'information. Cette notion d'accroissement local de l'entropie a été étudiée largement par CHUA sous la forme du *Local Activity Principle*, qui est introduit comme un troisième principe de la thermodynamique, permettant d'expliquer par des arguments mathématiques l'auto-organisation pour une certaine classe de systèmes complexes typiquement impliquant des équations de réaction-diffusion [MAINZER et CHUA, [2013](#)]. La manière dont l'information est stockée et compressée est essentielle pour la vie, puisque l'ADN est bien un système de stockage d'information (bien loin d'être compris complètement). La complexité culturelle implique un stockage de l'information bien plus complexe et à différents niveaux, et des flux d'information relevant fortement des deux autres types de complexité. Les flux d'information sont essentiels pour l'auto-organisation dans un système multi-agent. Les comportements collectifs de poissons ou d'oiseau sont des exemples typiques utilisés pour illustrer l'émergence et font partie des cas d'école de systèmes complexes. On commence cependant seulement à comprendre comment ces flux structurent le système, et quels sont les motifs spatiaux de transfert d'information au sein d'un *flock* par exemple : [CROSATO et al., [2017](#)] introduit des premiers résultats empiriques avec l'entropie de transfert pour des poissons et pose les bases méthodologiques de ce type d'étude.

PRODUCTION DE CONNAISSANCES Nous avons à présent la matière suffisante pour en venir à la réflexivité. Il est possible de positionner la production de connaissances à l'intersection des interactions entre types de complexité développées ci-dessus. Tout d'abord, la connaissance telle que nous l'envisageons ne peut se passer d'une construction collective, et implique donc un encodage et une transmission de l'information : il s'agit à un autre niveau de toutes les problématiques liées à la communication scientifique. La production de connaissances nécessite donc cette première interaction entre complexité computationnelle et complexité informationnelle. Le lien entre

complexité informationnelle et émergence est mobilisé si on considère l'établissement de connaissances comme un processus morphogénétique. Il est montré en 6.1 que le lien entre forme et fonction est fondamental en psychologie : nous pouvons l'interpréter comme un lien entre information et sens, puisque la sémantique d'un objet cognitif ne peut se passer d'une fonction. HOFSTADER rappelle dans [HOFSTADER, 1980] l'importance des symboles à différents niveaux pour l'émergence d'une pensée, qui consistent à un niveau intermédiaire en des signaux. Enfin, la dernière relation entre complexité computationnelle et émergence est celle qui nous permet d'affirmer qu'on s'intéresse particulièrement à une production de connaissance sur des systèmes complexes, les deux premiers pouvant s'appliquer à tout type de connaissance. Comme ces systèmes sont généralement multi-niveaux, ou présentent au moins un certain niveau de complexité computationnelle, leur connaissance se doit de la capturer, puisque même des modèles *simples* devront capturer leur complexité de manière conceptuelle et impliquer une structure conceptuelle sous-jacente complexe, même si celle-ci n'est pas explicitement explorée. Ainsi, toute connaissance complexe, ou *pensée complexe*, embrasse non seulement toutes les complexités mais aussi leur relations, dans son contenu et dans sa nature : elle doit nécessairement avoir un certain degré de réflexivité pour alors être cohérente. On peut tenter d'étendre à la réflexivité en tant que réflexion sur le positionnement disciplinaire : suivant PUMAIN dans [PUMAIN, 2005], la complexité d'une approche est également liée à la diversité des points de vue nécessaire pour la construire. Pour atteindre ce nouveau type de complexité, qui serait une dimension supplémentaire liée à la connaissance des systèmes complexes, la réflexivité doit être au coeur de la démarche. [READ, LANE et LEEUW, 2009] rappelle que l'innovation a été rendue possible quand les sociétés ont été capable de produire et diffuser de l'information sur leur propre structure, c'est à dire quand elles ont pu atteindre un certain niveau de réflexivité. La connaissance complexe serait donc le produit et le support de sa propre évolution grâce à la réflexivité qui a joué un rôle fondamental dans l'évolution du système cognitif : on pourrait ainsi suggérer de rassembler ces considérations, comme proposé par PUMAIN, sous une nouvelle notion épistémologique de *Rationalité Evolutive*.

* * *

*

CONCLUSION DU CHAPITRE

La lecture d'un article ou d'un ouvrage est toujours bien plus éclairante lorsqu'on connaît personnellement l'auteur, d'une part car on peut profiter des *private joke* et extrapoler certains développements des narrations qui se doivent synthétique (même si l'art de l'écriture est justement d'essayer de transmettre la majorité de ces éléments, l'ambiance en quelque sorte), et d'autre part car la personnalité a des implications complexes sur la manière d'appréhender la nature de la connaissance et une certaine structure a priori du monde. Pour cela, la connaissance scientifique serait très probablement moins riche si elle était produite par des machines aux capacités cognitives équivalentes, aux connaissances et expériences empiriques subjectives équivalentes et aussi diverses que celles humaines, mais qui auraient été programmées pour minimiser l'impact de leur personnalité et de leur convictions sur l'écriture et la communication (toujours en supposant qu'elles aient une certaine forme de données et fonctions plus ou moins équivalentes). Dans ces laboratoires de recherche dignes de *Blade Runner*, nous doutons que la production d'une pensée complexe serait effectivement possible, puisqu'il manquerait à ces machines justement la *Rationalité Evolutive* développée en 3.3, et nous doutons fortement que celle-ci puisse être produite du moins dans l'état des connaissances actuelles en intelligence artificielle. Le but de ce chapitre était donc "de faire connaissance" sur les points de positionnements incontournables pour l'ensemble de notre réflexion. Ceux-ci en sont d'autant plus en rien superflus car conditionnent très fortement certaines directions de recherche. Notre positionnement sur la reproductibilité développé en 3.1 implique certains choix de modélisation, notamment l'utilisation univoque de plateformes ouvertes, de workflow et d'implémentations ouverts ; il implique aussi un choix de données qui se doivent au maximum d'être accessibles ou rendues accessibles, et donc certains d'objets et d'ontologie, ou plutôt le non-choix de certains : nos problématiques pourraient être mobilisées sur des données d'entreprise fines tout en gardant une cohérence avec l'approche théorique et thématique (la théorie évolutive a largement mobilisé ce type d'étude comme par exemple [PAULUS, 2004]), mais la relative fermeture de ce type de données ne les rend pas utilisables dans notre démarche. Ensuite, notre positionnement sur le rôle du calcul intensif et les besoins d'exploration des modèles 3.2 est source de l'ensemble des expériences numériques et des méthodologies utilisées ou développées. Enfin, notre positionnement épistémologique 3.3 percole dans l'ensemble de notre travail, et permet de poser les premières briques pour des formalisations théoriques plus systématiques qui seront développées en Chapitre 9.

Deuxième partie

BRIQUES ÉLÉMENTAIRES

This part provides building blocks for the final objective of constructing models of co-evolution. These contain both stylized facts from empirical analyses and toy and hybrid modeling. They correspond to three distinct components of our overall construction : first analyses at the micro-scale confirming the chaotic and non-stationary nature of interactions between networks and territories, secondly a morphogenetic vision of these that corresponds roughly to a meso-scale, and finally an application of the evolutive urban theory at the macro-scale.

4

THÉORIE EVOLUTIVE URBAINE

TODO : quel niveau de description de la théorie évolutive ici ? insérer l'étude méthodes mixtes en dernier opening : illustration/construction du cadre de connaissances.

4.1 CORRÉLATIONS STATIQUES ENTRE FORME URBAINE ET FORME DE RÉSEAU

Study of interactions between network and territories :

→ *searching for stylized facts, what can be learnt from static correlations between urban form and road network?*

Theoretical Background : *A Theory of co-evolutive networked human territories* proposed in [RAIMBAULT, 2016g], that in particular postulates an important role of networks in the morphogenesis of complex adaptive urban systems that are human territories

→ *investigation of stationarity and ergodicity properties of relation between road network and population distribution ; implies spatiality of correlations and link static-dynamic*

C : (Florent) c'est trop technique comme entrée en matière ; pourquoi faudrait il qu'il y ait de la diffusion ? et de quels processus parles tu ? la forme urbain / réseaux / les deux ?

Spatio-temporal processes implying diffusion or propagation phenomena generally have a specific structure of correlation. In particular, as derived in section B.3, a static computation of correlation between different instances of a system may under certain conditions provide information on dynamical correlations implied.

4.1.1 Mesures morphologiques

Contexte

A l'échelle macroscopique du système de ville, le caractère spatial du système urbain est capturé de manière raisonnable par les positions des villes, associées aux variables agrégées au niveau de la ville qui représentent entièrement le système, comme la plupart des modèles liés à la Théorie Evolutive postulent. A l'échelle mesoscopique, à laquelle nous nous attendons à capturer des manifestations morphologiques des interactions entre ville et transport, la structure du système territorial peut être spécifiée par des indicateurs plus raffinés pour l'aspect morphologique. Le choix des indicateurs de forme urbaine pertinents pour répondre à un type de question donnée n'est pas évident, et dépendra de l'échelle et du contexte : on peut par exemple s'intéresser au caractère polycentrique pour lequel les indicateurs seront différents si on s'intéresse à des phénomènes de concentration. Notre but est de capturer le maximum de dimensions de variation de la forme urbaine, nous calculerons pour cela un certain nombre d'indicateurs arbitraire satisfaisant une certaine convergence de la variance cumulée des composantes principales.

FIGURE 8 :

FIGURE 9 : **C** : (Florent) attention au positionnement non controlé des figures en latex parfois malheureux comme ici **A1** : pb with page title ? add mdframed ?

Analyse Empirique

We study systematically morphological indicators for constant size areas covering European Community. The choice of fixed size areas can be questioned regarding definition of a territorial system, that can be otherwise understood as a consistent spatial entity at a given scale and along certain criteria : *Human territories* as defined by Raffestin (op. cit.) or more generally functionally autonomous spaces¹. Here we choose the mesoscopic scale of a metropolitan center ($\simeq 50\text{km}$) for comparability purposes and because greater scale are no more relevant regarding urban form, whereas smaller scales must contain too much noise.

C : (Florent) tu ne peux pas aller aussi vite sans choquer tout géographe quanti ; le fait de prendre des carrés 50×50 est arbitraire et si cela peut se justifier, il ne faut pas prétendre que tu as là des territoires comparables

Le but n'étant pas de comparer les territoires sur lesquels ces indicateurs sont calculés entre eux, mais de calculer une valeur "locale" et d'établir un champ discret régulier dans l'espace, la taille fixe de la fenêtre est nécessaire.

4.1.2 Mesures de Réseau

C : (Florent) et entre les deux il y a donc les indicateurs d'accessibilité que tu as évacué trop vite je pense (d'autant qu'étant pile à l'interface forme urbaine/réseau - ils me semblent particulièrement indiqués vu ta problématique

TODO : redo these analyses with accessibility indicators

TODO : recompute indicators with capacity and/or hierarchy when possible with speed limits, to check how they change, and also correlations.

¹ for example, a tentative of definition of a *Parisian* territory would present many facets. From the subjective territory point of view, intra-muros Parisians consider a strict boundary at *Boulevard périphérique*, **C** : (Florent) attention à bien intégrer les travaux des géographes sur cette épineuse Q cf Guéris Paulus 2002 [GUÉRIS et PAULUS, 2002] whereas close and even further suburbs will be seen as Parisians from the Province. The functional territory of *Metropolitain* extends slightly further than the administrative boundary. **C** : (Florent)laquelle la région, pas Paris Governance perimeters are currently mutating with the Metropolitan governance project. Complementary perceptions of the territory can thus be multiplied.

Nous considérons les mesures agrégées de réseau comme un moyen de caractériser les propriétés des réseaux de transport sur un territoire donné, de la même façon que les indicateurs morphologiques informent sur la structure urbaine. Nous proposons de calculer des indicateurs simples sur des étendues spatiales similaires à la morphologie, pour être en mesure d'explorer les relations entre ces mesures statiques. L'analyse statique de réseau a été intensément documentée dans la littérature, voir par exemple [LOUF et BARTHELEMY, 2014a] pour une étude comparative des villes ou [LAGESSE, 2015] pour l'exploration de nouvelles mesures pour le réseau de rues.

Pré-traitement des données

Nous travaillons ici avec le réseau de rues, dont la structure est finement conditionnée aux configurations territoriales des densités de population. De plus, les données du réseau de routes actuel est disponible ouvertement par l'intermédiaire du projet OpenStreetMap (OSM) [OPENSTREETMAP, 2012]. Sa qualité a été étudiée pour différents pays comme l'Angleterre [HAKLAY, 2010] et la France [GIRRES et TOUYA, 2010]. Il a été établi pour ces pays une qualité équivalente aux données officielles pour le réseau de rues primaire.

Pour les segments de rue primaires, nous calculons le réseau topologique pour l'ensemble des zones étudiées, à une granularité de 100m pour pouvoir être utilisé de manière cohérente avec les grilles de population. Les données OSM sont importées dans pgsql en utilisant osmosis [TEAM, 2016], est ensuite agrégé à la granularité fixe, et le réseau topologique résultant est finalement simplifié avec un algorithme split/merge.

$$\begin{aligned} &\simeq 44 \cdot 10^6 \text{ links in initial OSM db, } \simeq 61 \cdot 10^6 \text{ in first simplified layer,} \\ &\simeq 21 \cdot 10^6 \text{ in final database} \end{aligned}$$

ALGORITHME DE SIMPLIFICATION For a given dataset corresponding to a subset of the overall road network, it is necessary to simplify network structure by spatial aggregation as initial data presents very detailed features and thus a very large numbers of nodes ($\simeq 10^{10}$ for Europe dataset). **C : (Florent) c'est un peu confus, tu devrais d'abord dire : ce que tu as, les pb que ça pose, comment les résoudre** Such a level of precision is not needed in our study since density data is already aggregated at 500m resolution. It is possible to drastically reduce network size by spatial aggregation of nodes and link replacements. More precisely we use the following procedure :

- a background raster (which resolution r gives the snapping parameter for aggregation) is constructed from a reference raster and the extent of network. This grid gives spatial aggregation units for network nodes.

- for each feature of the road dataset, corresponding connected raster cells are stored with corresponding impedance and distance in a sparse adjacency matrix.
- Network is simplified by iterative suppression of nodes with degree two, with keeping link speed and real length to their effective value.

IMPLÉMENTATION A PostGIS database is used to store raw and simplified network, in order to perform efficient spatial requests, compared for example to initial osm data formats (osm or pbf). However the size of storage of data into this base is much higher (factor 10) so processing was parallelized between european countries. Consistence is ensured by the use of the same common density raster as simplification canvas. Final network is stored into the Postgis database for efficient indicator computation given a spatial extent. **C : (Florent)** *y'a t'il un effet de bord dans les carrés 50x50 qui se trouvent à la frontière de 2 pays* **A1 :** pas avec nouvelle parallelisation pas par pays mais par split and merge (TODO rewrite nouvel algo)

SENSIBILITÉ AUX PARAMÈTRES DE SIMPLIFICATION Sensitivity of indicators to raster resolution and to degree simplification algorithm must still be tested to ensure the relevance of data preprocesing.

Indicateurs

Network macroscopic structure is summarized by the following set of indicators, after the simplifications and reductions done in the previous step. Assuming network given by $N = (V, E)$, nodes spatial positions $\vec{x}(V)$ and edges *effective distances* $d(E)$ taking into account impedances and real distances (to include basically network hierarchy), we have indicators :

C : (Florent) tb à présenter de la même manière, plus en même temps + justification pour la forme urbaine

- connectivity
- degree distribution
- centrality, taken as normalized mean *betweenness-centrality*
- average path length
- network diameter
- mean network speed

These indicators are used to capture a rough picture of the structure. Refined work at smaller scales (intra-urban road network) and

with more elaborated measures that allow to differentiate more precisely local form, was recently done by Lagesse in [LAGESSE, 2015].

Résultats

Les indicateurs de réseau ont été calculés sur des zones similaires aux indicateurs de forme urbaine,

4.1.3 Correlations Statiques Effectives et Non-stationnarité

Corrélations spatiales

Results : Spatial Correlations

Computation of spatial correlation on square areas of width $\delta \cdot l_0$ (with typically $\delta = 4, \dots, 16$)

→ local spatial stationarity of processes

Results : Multi-scale Processes

→ Significant variation of mean correlation with δ (Left) and of normalized confidence interval (Right) given by $|\rho_+ - \rho_-| \cdot \delta$, as bounds theoretically vary as $\sqrt{N} \sim \sqrt{\delta^2}$: implies multi-scalarity

Application à la Chine

In [STEVENS et al., 2015] density grids for other countries across the world (ex. China) are provided² so we may repeat our analysis to other regions for comparison purposes. **C : (Florent) comparer quoi ?**

Non-stationnarité spatiale et non-ergodicité

Results : Computation of Indicators

Computation of urban form indicators [LE NÉCHET, 2015] and network indicators on $l_0 = 10\text{km}$ side square

Empirical Findings (Formalization)

$Y_i[\vec{x}, t]$ spatio-temporal stochastic process, verifies empirically :

1. Local spatial autocorrelation is present and bounded by l_ρ (in other words the processes are continuous in space) : at any \vec{x} and t , $|\rho_{\|\Delta\vec{x}\| < l_\rho} [Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$.
2. Processes are locally parametrized : $Y_i = Y_i[\alpha_i]$, where $\alpha_i(\vec{x})$ varies with l_α , with $l_\alpha \gg l_\rho$ and weakly locally stationary in space.

² available at <http://www.worldpop.org.uk/>

3. Processes are multi-scalar : since $\rho(\delta = \infty) > \rho(\delta = 0)$, a necessary non-linear correction on processes spatial averages in correlation computation is present.

Analytical Deductions

1. Regimes of temporal correlations. Let assume local ergodicity in \vec{x}_0 at scale $\delta \cdot l_0$ (reasonable with urban growth and network extension in recent times). The Ergodic theorem implies that $\exists \mathcal{T}$ such that

$$\langle Y_i(t) \rangle_{\|\vec{x} - \vec{x}_0\| < \delta \cdot l_0} = \langle Y_i(\vec{x}_0) \rangle_{t \in \mathcal{T}}$$

With spatial stationarity, $\langle Y_i \rangle_{\vec{x}_0} = \langle Y_i \rangle_{\vec{x}_1}$, thus \mathcal{T} must be constant to be invariant by translation. By contraposition and (2), processes have different dynamical characteristics.

2. Global non-ergodicity. Let X_k a partition of space into local areas. We have $\langle \cdot \rangle_x = \sum_k w_k \langle \cdot \rangle_{x_k} = (1) \sum_k w_k \langle \cdot \rangle_{\mathcal{T}_k}$. On the other hand, global ergodicity would give $\langle \cdot \rangle_t = \langle \cdot \rangle_{\mathcal{T}} = \sum_k w_k \langle \cdot \rangle_{\mathcal{T}}$ and $\sum_k w_k (\langle \cdot \rangle_{\mathcal{T}} - \langle \cdot \rangle_{\mathcal{T}_k}) = 0$. Being true on each subset implies $\mathcal{T} = \mathcal{T}_k$, what contradicts (1).

Case study : implications

→ Still points to explore :

- variable correlations areas (size and shape in space)
- same work on cities population/train network data, which are also dynamical databases : extrapolation of ergodicity parameters ?
- correlations of returns : link between $\rho[\Delta_t Y]$ and $\rho[\Delta_x Y]$ (more difficult : if pure local ergodicity, \exists a permutation making the correspondance)
- Link between $\Delta_\delta \rho(\delta)$ and process derivatives ?

→ We show the regional nature of network-territories interactions, in particular the non-ergodicity of urban systems on **the interaction these components**

→ No direct results on time dynamics, but indirect : spatio-temporal processes do not have same speed and react/diffuse differently

4.2 CAUSALITÉS SPATIO-TEMPORELLES

Spatial statistics studies on dynamical relations between network and territories are relatively rare. [LEVINSON, 2008] does so on London metropolitan area and identifies causalities using lagged variables, but does not disentangle relations in the sense of coupled statistical models that would isolate endogenous effects from coupling effects.

Formalisation

We assume a dynamic transportation network $n(\vec{x}, t)$ within a dynamic territorial landscape $\vec{T}(\vec{x}, t)$, which components are to simplify population $p(\vec{x}, t)$ and employments $e(\vec{x}, t)$. Data is structured the following way :

- Observation of territorial variables are discretized in space and in time, i.e. the spatial field \vec{T} is summarized by $T = (\vec{T}(\vec{x}_i, t_j^{(T)}))_{i,j}$ with $1 \leq i \leq N$ and $1 \leq j \leq T$. They concretely correspond to census on administrative units (*communes* in our case) at different dates.
- Network has a continuous spatial position but is represented by the vector of network distances N **C : (Florent) vol d'oiseau/-distance temps ? second faisable et à privilégier je pense**

Sur l'accessibilité

The notion of accessibility has been central to regional science since its introduction and systematization in planning around 1970.

As already introduced in the first chapter, we question the notion of accessibility : *Is the notion of accessibility crucial for statistical analysis ?*

Weibull has proposed an axiomatic approach to accessibility [WEIBULL, 1976], deriving a canonical decomposition for any *attraction-accessibility* function $A(a, d)$, assuming expected thematic axioms among others technical ones that are :

1. A is invariant regarding the order of the configuration
2. A decrease with distance at fixed attraction and increase with attraction at fixed distance
3. A is invariant when adding null attractions and constant configurations

Then A verifies these if and only if it is of the form

$$A[(a_i, d_i)] = T \left(\bigoplus_i z(d_i, a_i) \right)$$

where T is increasing with null origin, z is a *distance substitution function* (i.e. verifying axiom 2) and \oplus a *standard composition* associating two attractions at zero distance to the corresponding unique one.

It means that well suited matrices of autocorrelation should capture accessibility in regressions; **C : (Florent) pas sur de comprendre, à discuter** or it must be captured by non-linear regression on N . It may reveal some kind of intrinsic accessibility that is related to real phenomena (that we expect to fit with calibrated functions of accessibility based on Hedonic models e.g.) Seeing accessibility as a potential field is an equivalent vision : given any stationary dynamic for n, \vec{T} , Helmholtz theorem states that it derives from a potential (can be adapted to non-stationary dynamics with a time-varying potential).

Données

We will work on a novel dataset provided by LE NECHET, that consists in main road infrastructures with their opening dates and train network for network dynamics, and in population and employments of communes at census dates, for Bassin Parisien on the last fifty year. The temporal granularity due to census temporal step may be an obstacle to obtain good dynamical statistics. **C : (Florent) enfin c'est surtout INSEE, IGN, et Wiki[?] qu'il faut citer (c'est vrai qu'il y a du formatage, mais en tout cas il faut citer les sources de première main)**

Tests Statistiques

The following large set of analysis are to be tested (non exhaustive) :

C : (Florent) interprétation ? si O/N

- On raw data :

- Multivariate models

$$\mathcal{L} [\mathbf{T}, \mathbf{N}] \sim \varepsilon$$

- Autocorrelated univariate models

$$(\mathbf{I} - \Sigma \mathbf{R} \mathbf{W}) \mathbf{X} \sim \varepsilon$$

- Autocorrelated multivariate models

$$(\mathcal{L}' - \Sigma \mathbf{R} \mathbf{W}) [\mathbf{T} + \mathbf{N}] \sim \varepsilon$$

- Geographically Weighted Regression [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998]

$$\mathcal{L} [\mathcal{G} (\mathbf{T}, \mathbf{N})] \sim \varepsilon$$

- Granger causality tests : [XIE et LEVINSON, 2009a] use for example Granger causality to link transit with land-use changes.
- On data returns :
 - Autoregressive multivariate models

$$\mathcal{L} [(\Delta T(t_{j'}))_{j' \leq j}, (\Delta N(t_{j'}))_{j' \leq j}] \sim \varepsilon$$

- Autoregressive autocorrelated multivariate models : idem with spatial autocorrelation term.
- Synthetic Instrumental Variables : static territory and/or network?

4.2.1 Une méthode pour identifier des causalités spatio-temporelles

L'étude des processus spatio-temporels fortement couplés implique la prise en compte d'intrications entre ceux-ci généralement difficiles à isoler. Essence même des approches par la complexité, ces interactions qui sont à l'origine du comportement émergent d'un système font sens comme objet d'étude en lui-même, et une séparation des processus paraît alors contradictoire avec une vision intégrée du système. Dans le cas des systèmes territoriaux, l'exemple des interactions entre réseaux de transport et territoires est une excellente allégorie de ce phénomène : des méthodes isolant les "effets structurants" d'une infrastructure développées dans les années 70 [BONNAFOUS et PLASSARD, 1974] se sont révélées par la suite de l'instrumentation politique et sans fondement empirique [OFFNER, 1993]. Le débat est toujours d'actualité puisque la question se pose toujours par exemple pour la construction de lignes à grande vitesse [CROZET et DUMONT, 2011b]. La réalité des processus territoriaux est en fait bien plus compliquée qu'une simple relation causale entre la mise en place d'une infrastructure et les retombées sur le développement local, mais correspond bien d'une *co-évolution* complexe [BRETAGNOLLE, 2009b]. Sur le temps long et à grande échelle, certains effets de renforcement des dynamiques dans les systèmes de villes par l'insertion dans les réseaux, ont été mis en valeur par l'application de la Théorie Evolutive des Villes [L'ESPACE GÉOGRAPHIQUE, 2014], montrant que le démêlage est toutefois possible dans certains cas par une compréhension plus globale du système. A une autre échelle, toujours concernant les relations entre réseaux et territoires, on peut citer les liens entre pratiques de mobilité, également urbain et localisation des ressources dans un cadre métropolitain [CERQUEIRA, 2017] qui s'avèrent tout autant complexes. Ce type de problématique est bien sûr présent dans d'autres domaines : en Economie Géographique, l'exemple des liens

entre innovation, impacts locaux de la connaissance et aggregation des agents économiques est une illustration typique de processus économiques spatio-temporels présentant des causalités circulaires difficiles à démêler [AUDRETSCH et FELDMAN, 1996]. Des méthodes spécifiques sont introduites, comme l'utilisation d'instruments statistiques comme par [AGHION et al., 2015] dans lequel l'origine géographique des membres du Bureau du Congrès américain attribuant les subventions locales est une bonne variable instrumentale pour lier caractère innovant et inégalités des plus haut salaires, et permet de montrer que la corrélation significative entre les deux est en fait une causalité de l'innovation sur les inégalités.

Le couplage fort spatio-temporel implique généralement l'introduction de la notion de causalité, à laquelle la géographie s'est toujours intéressée : [LOI, 1985] montre que les questions fondamentales que se pose la géographie théorique récente (isolation des objets, lien entre espace et structures causales, etc.) étaient déjà présentes dans la géographie classique de Vidal. [CLAVAL, 1985] critique d'ailleurs les nouveaux déterminismes ayant émergé, notamment celui proposé par certains tenants de l'analyse systémique : dans ses débuts, cette approche héritait de la cybernétique et donc d'une vision réductionniste impliquant un déterminisme même dans une formulation probabiliste. Claval note que des travaux contemporains à son écriture devraient permettre de capturer la complexité qui fait la particularité des décisions humaines : l'école de Prigogine et la Théorie des Catastrophes de Thom. Ce point de vue est remarquablement visionnaire, puisque comme le rappelle Pumain dans [PUMAIN, 2003], le glissement de l'analyse des systèmes à l'auto-organisation puis à la complexité a été long et progressif, et ces travaux ont été fondamentaux pour le permettre. François Durand-Dastès résume cette situation plus récemment dans [DURAND-DASTES, 2003], en appuyant l'importance des bifurcations et de la dépendance au chemin lors des instants initiaux de la constitution du système qu'il désigne par *systèmogenèse*. Ce type de dynamique complexe implique généralement une co-évolution des composantes du système, qu'on peut interpréter comme des causalités circulaires entre processus : la question de pouvoir les identifier est donc cruciale au regard de la notion de causalité pour la géographie complexe contemporaine.

Les régimes sous lesquels des identifications de causalité sont cohérentes ne sont pas identifiés de manière évidente. Ceux-ci dépendront des définitions utilisées, de la même manière que les méthodes à disposition pour lesquelles nous pouvons donner quelques illustrations. [LIU et al., 2011] propose la détection de relations spatio-temporelles entre perturbations des flots de trafic, introduisant une définition particulière de la causalité basée sur une correspondance de points extrêmes. Les algorithmes associés sont toutefois spécifiques et difficilement applicables à des types de systèmes différents. L'uti-

lisation des correlations spatio-temporelles a été démontrée comme ayant dans certains cas un fort pouvoir prédictif pour les flots de traffic [MIN et WYNTER, 2011]. Également dans le domaine des transports et de l'usage du sol, [XIE et LEVINSON, 2009a] applique une analyse par causalité de Granger, qu'on pourra interpréter comme une corrélation retardée, pour montrer dans un cas particulier que la croissance du réseau induit le développement urbain et est elle-même tirée par des externalités comme les habitudes de mobilité. Les neurosciences ont développé de nombreuses méthodes répondant à des problématiques similaires. [LUO et al., 2013] définit une causalité de Granger généralisée prenant en compte la non-stationnarité et s'appliquant à des régions abstraites issues d'imagerie fonctionnelle. Ce genre de méthode est également développée en Vision par Ordinateur, comme l'illustre [KE, SUKTHANKAR et HEBERT, 2007] qui exploite les correlations spatio-temporelles de formes et de flux dans des successions d'images pour classifier et reconnaître des actions. Les applications peuvent être très concrète comme la compression de fichier vidéos par extrapolation des vecteurs de mouvement [CHALIDABHONGSE et KUO, 1997]. Dans l'ensemble de ces cas, l'étude des correlations spatio-temporelles rejoint les notions faibles de causalité vues précédemment. Cette contribution cherche à explorer la possibilité d'une méthode analogue pour des données spatio-temporelles présentant a priori des causalités circulaires complexes, et donc de tenter l'exercice d'équilibrisme de concilier un certain niveau de simplicité et de caractère opérationnel à une prise en compte de la complexité. Nous introduisons ainsi une méthode d'analyse des correlations spatio-temporelles similaire à une causalité de Granger estimée dans le temps et l'espace, dont la robustesse est démontrée systématiquement par l'application à un modèle de simulation complexe de morphogenèse urbaine et par l'isolation de régimes de causalités distincts dans l'espace des phases du modèle. Notre contribution inclut également l'application à un cas d'étude empirique, ce qui la positionne à l'interface des domaines de la méthodologie, de la modélisation et de l'empirique.

La suite de cette section est organisée de la façon suivante : le cadre générique de la méthode proposée est décrit. Nous l'appliquons ensuite à un jeu de données synthétiques afin de la valider partiellement et de tester ses potentialités, ce qui permet de l'appliquer ensuite au système urbain Sud-Africain sur le temps long. Nous discutons finalement la proximité avec d'autres méthodes existantes et des développements possibles.

Méthode

Nous formalisons ici de manière générique la méthode, basée sur un test similaire à la causalité de Granger, pour tenter d'identifier des relations causales dans des systèmes spatiaux. Soit $X_j(\vec{x}, t)$ des proces-

sus aléatoires spatiaux unidimensionnels, se réalisant dans le temps et l'espace. On se donne un ensemble d'unités spatiales fondamentales (u_i) qui peuvent être par exemple les cellules d'un raster ou un pavage quelconque de l'espace géographique. On suppose l'existence de fonctions $\Phi_{i,j}$ permettant de faire correspondre les réalisations de chaque composante aux unités spatiales, possiblement par une première agrégation locale. Une réalisation d'un système est donnée par un ensemble de trajectoires pour chaque processus $x_{i,j,t}$, et on pourra noter un ensemble de réalisations $x_{i,j,t}^{(k)}$ (accessibles dans le cas d'un modèle de simulation par exemple, ou par hypothèse de comparabilité de sous-systèmes territoriaux dans des cas réels). On suppose disposer d'un estimateur de corrélation $\hat{\rho}$ s'exerçant dans le temps, l'espace et les répétitions, i.e. $\hat{\rho}[X, Y] = \hat{E}_{i,t,k}[XY] - \hat{E}_{i,t,k}[X]\hat{E}_{i,t,k}[Y]$. Il est important de noter ici l'hypothèse de stationnarité spatiale et temporelle, qui peut toutefois aisément se relâcher dans le cas d'une stationnarité locale. D'autre part, l'autocorrelation spatiale n'est pas explicitement incluse, mais est prise en compte soit par l'agrégation initiale si l'échelle caractéristique des unités est plus grande que celle des effets de voisinage, soit par un estimateur spatial adéquat (statistiques spatiales pondérées de type GWR [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998] par exemple). Cela nous permet de définir la corrélation retardée par

$$\rho_\tau[X_{j_1}, X_{j_2}] = \hat{\rho}\left[x_{i,j_1,t-\tau}^{(k)}, x_{i,j_2,t}^{(k)}\right] \quad (2)$$

La corrélation retardée n'est pas directement symétrique, mais on a de manière évidente $\rho_\tau[X_{j_1}, X_{j_2}] = \rho_{-\tau}[X_{j_2}, X_{j_1}]$. On applique alors cette mesure de manière simple : si $\text{argmax}_\tau \rho_\tau[X_{j_1}, X_{j_2}]$ ou $\text{argmin}_\tau \rho_\tau[X_{j_1}, X_{j_2}]$ sont "clairement définis" (les deux pouvant l'être simultanément), leur signe donnera alors le sens de la causalité entre les composantes j_1 et j_2 et leur valeur absolue le retard de propagation. Les critères de significativité dépendront du cas d'application et de l'estimateur utilisé, mais peuvent par exemple inclure la significativité du test statistique (test de Fisher dans le cas d'un estimateur de Pearson), la position des bornes d'un intervalle de confiance à un niveau donné, ou même un seuil exogène θ sur $|\rho_\tau|$ pour forcer un certain degré de corrélation.

4.2.2 Données Synthétiques

Cette méthode doit dans un premier temps être testée et partiellement validée, ce que nous proposons de faire sur des données synthétiques, méthode qui permet une connaissance plus fine des comportements des modèles [RAIMBAULT, 2016a]. En écho à l'exemple des relations entre réseaux de transport et territoires qui a permis d'introduire notre problématique précédemment, nous proposons de

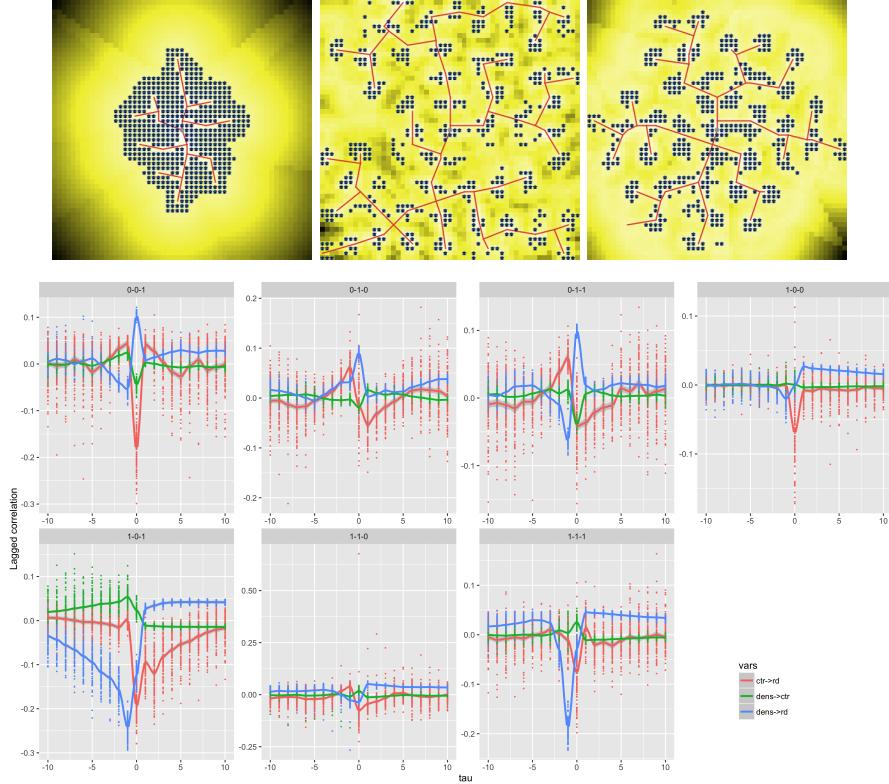


FIGURE 10 : Correlations dans le modèle RDB (Première ligne) Exemples de configurations finales variées, obtenues avec (w_d, w_c, w_r) valant respectivement $(0, 1, 1), (1, 0, 1)$, et $(1, 1, 1)$. (Deuxième ligne) Corrélations retardées, pour chaque combinaison des paramètres, en fonction du retard τ . Les différentes couleurs correspondent à chaque couple de variables : distance au centre (*ctr*), densité (*dens*) et distance au réseau (*rd*). Les points montrent l'étendue sur l'ensemble des répétitions du modèle (estimateurs sur i et t).

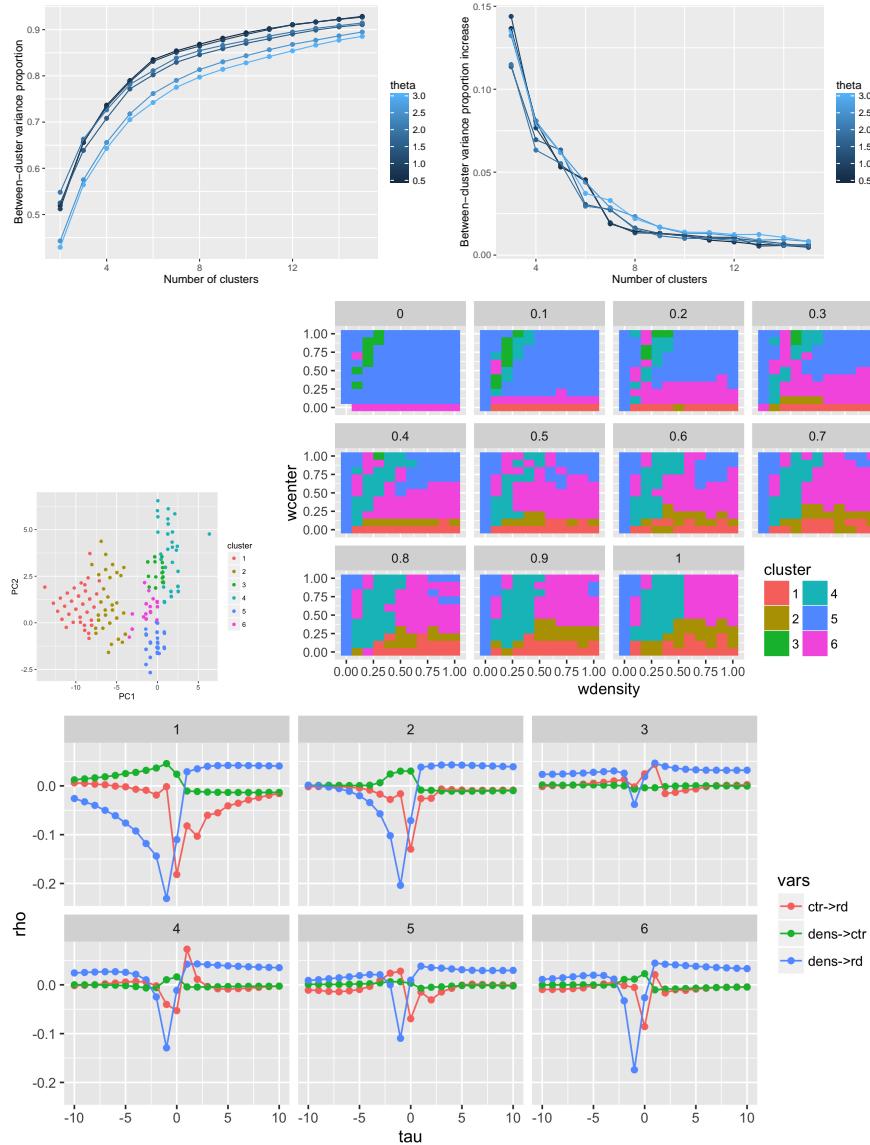


FIGURE 11 : Identification de régimes d’interactions **(Haut Gauche)** Variance inter-cluster comme fonction du nombre de clusters. **(Haut Droite)** Dérivée de la variance inter-cluster. **(Milieu Gauche)** Features dans un plan principal (81% de variance expliquée par les deux premières composantes) **(Milieu Droite)** Diagramme de phase des régimes dans l’espace (w_d, w_c, w_r), w_r variant entre les différents sous-diagrammes de (w_d, w_c). **(Bas)** Trajectoires correspondantes des centroïdes.

générer des configurations urbaines stylisées dans lesquelles réseau et densité s'influencent mutuellement, et pour lesquelles les causalités ne sont pas évidents *a priori* étant donné les paramètres du modèle génératif. [RAIMBAULT, BANOS et DOURSAT, 2014] décrit et explore un modèle simple de morphogenèse urbaine (modèle RBD) répondant parfaitement à ces contraintes. En effet, les variables explicatives de la croissance urbaine, les processus d'extension du réseau et le couplage entre densité urbaine et réseau ne sont pas trop complexes. Cependant, hormis dans des cas extrêmes (par exemple lorsque la distance au centre détermine la valeur foncière uniquement, le réseau dépendra de manière causale de la densité, ou lorsque la distance au réseau seule compte, la causalité sera inversée), les régimes mixtes n'exhibent pas de causalités évidentes : c'est donc un parfait cas pour tester si la méthode est capable d'en détecter. Nous utilisons une implémentation adaptée³ du modèle initial, permettant de capturer les valeurs des variables étudiées pour chaque patch et à chaque pas de temps et de calculer les corrélations retardées entre variables au sein du modèle. Nous explorons une grille de l'espace des paramètres du modèle RBD, faisant varier les paramètres de poids de la densité, de la distance au centre et de la distance au réseau⁴, que l'on note respectivement (w_d, w_c, w_r), dans $[0; 1]$ avec un pas de 0.1. Les autres paramètres sont fixés à leur valeurs par défaut données par [RAIMBAULT, BANOS et DOURSAT, 2014]. Pour chaque valeur des paramètres, nous procédons à $N = 100$ répétitions ce qui est suffisant pour une bonne convergence des indicateurs. Les explorations sont effectuées via le logiciel OpenMole [REUILLOU, LECLAIRE et REY-COYREHOURCQ, 2013], le grand nombre de simulations (1,330,000) nécessitant l'utilisation d'une grille de calcul. Nous calculons sur l'ensemble des patches les corrélations retardées par estimateur de Pearson non biaisé entre les variations des variables suivantes⁵ : densité locale, distance au centre et distance au réseau. La Fig. 10 montre le comportement de ρ_τ pour chaque couple de variable (non dirigé, τ prenant des valeurs négatives et positives), pour les combinaisons des valeurs extrêmes des paramètres. On peut voir déjà différents régimes émerger : par exemple, (1, 0, 1) conduit à une causalité de la densité sur la distance au centre avec un retard 1, et une causalité négative de la densité sur la distance au réseau avec le même retard, tandis que distance au centre et au réseau sont corrélés de manière synchrone. Afin d'étudier ces comportements de manière systématique, nous proposons d'identifier des régimes de manière endogène, en procédant à un apprentissage.

³ disponible sur le dépôt ouvert du projet à
<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Simple/ModelCA>

⁴ Le modèle fonctionne de la façon suivante : une valeur des patches est déterminée par la moyenne pondérée de ces différentes variables explicatives, valeur qui détermine la croissance de nouveaux patches à l'instant suivant.

⁵ Calculer les corrélations sur les variables directement n'a pas de sens puisque leur valeur n'en a pas en absolu.

sage non-supervisé. Nous appliquons une classification des *k-means*, robuste à la stochasticité (5000 répétitions), avec les points caractéristiques (*features*) suivants : pour chaque couple de variable, $\text{argmax}_\tau \rho_\tau$ et $\text{argmin}_\tau \rho_\tau$ si la valeur correspondante est telle que $\frac{\rho_\tau - \bar{\rho}_\tau}{|\bar{\rho}_\tau|} > \theta$ avec θ paramètre de seuil, 0 sinon. L'inclusion des *features* supplémentaires des valeurs de ρ_τ n'influence pas significativement les résultats, celles-ci n'ont pas été prises en compte pour réduire la dimension. Le choix du nombre de clusters k est en général épineux dans ce genre de problème [HAMERLY et ELKAN, 2003], dans notre cas le système possède une structure agréable : les courbes de la proportion de variance inter-cluster et de sa dérivée en Fig. 11, en fonction de k pour différentes valeurs de θ , présentent une transition pour $\theta = 2$, ce qui donne pour cette courbe une rupture à $k = 5$. Un examen visuel des clusters dans un plan principal confirme la bonne qualité de la classification pour ces valeurs. Une classe correspond alors à un *régime de causalité*, dont nous pouvons représenter le diagramme de phase en fonction des paramètres du modèle, ainsi que les trajectoires des centres des clusters (calculées comme barycentre dans l'espace complet initial) en Fig. 11. Le comportement obtenu est particulièrement intéressant : les régions du diagramme correspondant aux régimes sont clairement délimitées et connexes. Par exemple, on observe l'émergence du régime 6 où la distance au réseau cause fortement la densité de manière négative, mais la distance au centre cause la distance au réseau, régime dont l'étendue maximale sur (w_d, w_r) est pour une valeur intermédiaire $w_r = 0.7$. Ainsi, pour maximiser l'impact du réseau sur la densité, il ne faut pas maximiser le poids correspondant, ce qui peut paraître contre-intuitif en premier abord : cela illustre l'intérêt de la méthode dans le cas de relations circulaires difficiles à démêler a priori. Le régime 5, où la distance au réseau influence la densité de la même manière, mais la relation entre distance au centre et route est inversée, est tout aussi intéressant, et est prédominant dans les faibles w_r . Le régime 1, extrême, correspond à une situation isolée dans laquelle la distance au centre n'importe pas : cet aspect domine alors totalement les autres processus d'interaction entre densité et réseau. Cette application sur données synthétique démontre ainsi d'une part la robustesse de la méthode vu la cohérence des régimes obtenus, et constitue aussi une qualification beaucoup plus précise des comportements du modèle que celle réalisée dans l'article initial. Dans ce cas précis, il peut s'agir d'un instrument de connaissance des relations entre réseaux et territoires en lui-même, permettant le test d'hypothèses ou la comparaison de processus dans le modèle stylisé.

4.2.3 Relations Réseaux-territoires en Afrique du Sud

Contexte

Transportation Networks can be leveraged as a powerful socio-economic control tool, with even more significant outcomes when it percolates to their interaction with territories. The case of South Africa is an accurate illustration, as [BAFFI, 2016] shows that during apartheid railway network planning was used as a racial segregation tool by shaping strongly constrained mobility and accessibility patterns. In particular, it is shown qualitatively that dynamics between territories and networks profoundly changed at the end of the apartheid, transforming a tool of planned segregation (network shaped was optimized to minimize unwanted accessibility) into an integration tool thanks to recent changes in network topology patterns.

We propose to investigate the potential *structural* properties of this historical process, by focusing on dynamical patterns of interactions between the railway network and city growth. More precisely, we try to establish if the segregative planning policies did actually modify the trajectory of the coupled system, what would correspond to deeper and wider impacts.

Objectifs

We can use first the particular shape of that network to control on local and global topology effects (but this is quite equivalent as controlling on accessibility), and in a second time the historical events as statistic instruments, assuming that territorial dynamics and network dynamics responded differently to these. We expect to learn from these project informations on interactions at long time scale and large spatial scale, in a very particular context of constrained growth. **C :**
(Florent) à discuter

Développements possibles

The method of instruments in statistics [ANGRIST, IMBENS et RUBIN, 1996] is used to identify causal relationships between variables, in a different way than Granger causality test for example. Trying to identify causalities between network dynamics and territorial dynamics is of crucial importance to test our theoretical assumption on the existence of co-evolution.

4.3 EFFETS DE RÉSEAU RÉVÉLÉS PAR UN MODÈLE DE CROISSANCE MACROSCOPIQUE

4.3.1 *Contexte*

4.3.2 *Modèle et Résultats*

4.3.3 *Discussion*

CONCLUSION DU CHAPITRE

5

ECHELLES ET ONTOLOGIES

La richesse des interactions entre réseaux et territoires, développée en Chapitre 1, est que celle-ci occurrent à différentes échelles, entre ces échelles, et par des intermédiaires très variés, au sens des agents ou structures impliquées mais aussi de leur caractéristiques, ceux-ci allant de la congestion des réseaux aux dynamiques sur le temps long en passant par les re-localisations des activités par exemple. Le cas de Zhuhai développé en 1.2 illustre la complexité d'une trajectoire locale et régionale, d'une bifurcation politique induisant l'instauration de la Zone Economique Spéciale par XI JINPING conditionnée à une bifurcation historique bien plus ancienne liée à la colonisation européenne qui a conduit à l'existence de Macao, à une bifurcation géographique en terme d'accessibilité régionale et une nouvelle position centrale de la ville dans la Mega-city Region du Delta de la Rivière des Perles. Nous avons dans le chapitre précédent étudié empiriquement les manifestations morphologiques des interactions à l'échelle mesoscopique, mais également mis en évidence des effets de structure à cette même échelle sur un temps long dans le cas de l'Afrique du Sud. Quelle échelle minimale est-il pertinent de considérer, autrement dit l'étude de l'échelle microscopique peut-elle nous apporter de l'information ? Et peut-on clarifier certaines ontologies, ou au moins un certain degré de précision ou de complexité requis dans celles-ci ? Ce chapitre cherche à répondre à ces interrogations par le biais d'études empiriques. Ainsi, nous tentons de préciser itérativement la structure des modèles futurs, mais aussi leur non-structure.

Dans une première section 5.1, nous explorons empiriquement un jeu de données à l'échelle microscopique sur le traffic routier en Ile-de-France, en ayant notamment à l'esprit la notion d'équilibre des flots de traffic qui est une hypothèse particulièrement répandue dans la modélisation du traffic. Nous démontrons que cet équilibre n'a aucun fondement empirique, et que les trajectoires microscopiques du système sont chaotiques. Cela nous permettra d'une part de confirmer nos choix épistémologique de modèle loin de l'équilibre typique d'une appréhension de la complexité, d'autre part de confirmer que cette échelle n'est pas pertinente. Nous continuons sur le traffic routier dans une deuxième section 5.2, en nous concentrerons sur la composante du prix de transport via le proxy du prix de vente du carburant, et ces liens potentiels avec les caractéristiques socio-économiques des territoires, dans le cas des Etats-Unis avec une granularité spatiale au Conté et temporelle à la journée. Nous obtenons le résultat assez inattendu des deux échelles endogènes proprement définies, corres-

pondant aux échelles mesoscopique et macroscopique, mais aussi la mise en évidence de la superposition de processus de gouvernance à des processus locaux. Enfin, la dernière section 5.3 applique la méthode d'identification de causalités développée en 4.2 au différents projets de transport du Grand Paris et démontre des potentiels effets d'annonce des projets de transport sur la croissance de la population, confirmant la pertinence d'une échelle d'agrégation au moins mesoscopique et de se concentrer sur des variables territoriales relativement basiques.

★ ★

★

Ce chapitre est entièrement adapté de divers articles : la section 5.1 a été publiée en anglais comme [RAIMBAULT, 2017f] ; la section 5.2 également en anglais en collaboration avec A. BERGEAUD comme [RAIMBAULT et BERGEAUD, 2017] ; la section 5.3 correspond à la partie d'application de [RAIMBAULT, 2017e].

5.1 INVESTIGATION EMPIRIQUE DE L'EXISTENCE DE L'EQUILIBRE UTILISATEUR STATIQUE

L'Equilibre Utilisateur Statique est un cadre puissant pour l'étude théorique du trafic. Malgré l'hypothèse restreignant de stationnarité des flots qui intuitivement limite son application aux systèmes de trafic réels, de nombreux modèles opérationnels qui l'implémentent sont toujours utilisés sans validation empirique de l'existence de l'équilibre. Nous étudions celle-ci sur un jeu de données de trafic couvrant trois mois sur la région parisienne. L'implémentation d'une application d'exploration interactive de données spatio-temporelles permet de formuler l'hypothèse d'une forte hétérogénéité spatiale et temporelle, guidant les études quantitatives. L'hypothèse de flots localement stationnaires est invalidée en première approximation par les résultats empiriques, comme le montrent une forte variabilité spatio-temporelle des plus courts chemins et des mesures topologiques du réseau comme la centralité de chemin. De plus, le comportement de l'index d'autocorrelation spatiale pour les motifs de congestion à différentes portées spatiales suggère une évolution chaotique à l'échelle locale, en particulier lors des heures de pointe. Nous discutons finalement les implications de ces résultats empiriques et proposons des possibles développements futurs basés sur l'estimation de la stabilité dynamique au sens de Lyapounov des flots de trafic.

5.1.1 Contexte

La modélisation du trafic a été largement étudiée depuis les travaux séminaux de Wardrop ([WARDROP, 1952]) : les enjeux économiques et techniques justifient entre autre le besoin d'une compréhension fine des mécanismes régissant les flots de trafic à différentes échelles. Différentes approches aux objectifs différents coexistent aujourd'hui, parmi lesquels on trouve par exemple les modèles dynamiques de micro-simulation, généralement opposés aux techniques de basant sur l'équilibre. Tandis que la validité des modèles microscopiques a été étudiée de façon conséquente et leur application souvent questionnée, la littérature est relativement pauvre en études empiriques assurant l'hypothèse d'équilibre stationnaire du cadre de l'Equilibre Utilisateur Statique (EUS). De nombreux développements plus réalistes on été documentés dans la littérature, tels l'Equilibre Utilisateur Dynamique Stochastique (EUDS) (voir pour une description par exemple [HAN, 2003]). A un niveau intermédiaire entre les cadres statiques et stochastiques se trouve l'Equilibre Utilisateur Stochastique Restreint, pour lequel les choix d'itinéraire des utilisateurs sont contraints à un ensemble d'alternatives réalisables ([RASMUSSEN et al., 2015]). D'autres extensions prenant en compte le comportement de l'utilisateur via des modèles de choix ont été proposé plus récem-

ment, comme [ZHANG, MAHMASSANI et LU, 2013] qui inclut à la fois l'influence de la tarification routière et de la congestion sur le choix avec un modèle Probit. La relaxation d'autres hypothèses restrictives comme la maximisation pure de l'utilité par l'utilisateur ont aussi été introduites, tels l'Equilibre Utilisateur Borné décrit par [MAHMASSANI et CHANG, 1987]. Dans ce cadre, l'utilisateur est satisfait si son utilité tombe dans un intervalle et l'équilibre est achevé lorsque chaque utilisateur est satisfait. Les dynamiques résultantes sont plus complexes comme révélé par l'existence d'équilibres multiples, et permet de rendre compte de faits stylisés spécifiques comme des évolutions irréversibles du réseau comme développé par [Guo et LIU, 2011]. D'autres modèles d'attribution de trafic inspirés d'autres domaines ont également été plus récemment proposés : dans [PUZIS et al., 2013], une définition étendue de la centralité de chemin qui combine linéairement la centralité des flots non-constraints avec une centralité pondérée par le temps de parcours permet d'obtenir une forte corrélation avec les flots de trafic effectifs, fournissant ainsi un modèle d'attribution de trafic. Cela fournit également des applications pratiques comme l'optimisation de la distribution spatiale des capteurs de trafic.

Malgré ces nombreux développements, de nombreuses études et applications concrètes se reposent toujours sur l'Equilibre Utilisateur Statique. La région parisienne utilise par exemple un modèle statique (MODUS) pour gérer et planifier le trafic. [LEURENT et BOUJNAH, 2014] introduit un modèle statique de flots qui inclut les recherches locales et le choix du parking : il est légitime de s'interroger, en particulier à de si faibles échelles, si la stationnarité de la distribution des flots est une réalité. Une example d'exploration empirique des hypothèses classiques est donné par [ZHU et LEVINSON, 2010], pour lequel les choix d'itinéraires révélés sont étudiés. Les conclusions questionnent le "premier principe de Wardrop" qui implique que les utilisateurs choisissent parmi un ensemble d'alternatives parfaitement connu. Dans le même esprit, nous étudions l'existence possible de l'équilibre en pratique. Plus précisément, l'EUS suppose une distribution stationnaire des flots sur l'ensemble du réseau. Cette hypothèse reste valable dans le cas d'une stationnarité locale, tant que l'échelle temporelle d'évolution des paramètres est considérablement plus grande que les échelles typiques de voyage. Le second cas qui est plus plausible et de plus compatible avec les cadres théoriques dynamiques est testé ici.

La suite de ce travail s'organise ainsi : la procédure de collection de données ainsi que le jeu de données sont décrits ; nous présentons ensuite une application interactive pour l'exploration du jeu de données, dans le but de fournir une intuitions sur les motifs présents ; puis nous donnons divers résultats d'analyses quantitatives allant dans le sens d'indices convergents pour une non-stationnarité des flots de

trafic; nous discutons finalement les implications de ces résultats et des développements possibles.

5.1.2 Résultats

Collecte des données

CONSTRUCTION DU JEU DE DONNÉES Nous proposons de travailler sur l'étude de cas de la région métropolitaine de Paris. Un jeu de données ouvert a été construit, comprenant les liens autoroutiers dans la région, par collecte des données publiques en temps réel des temps de parcours (disponible sur www.sytadin.fr). Comme rappelé par [BOUTEILLER et BERJOAN, 2013], la disponibilité de jeux de données ouverts pour les transports est loin d'être la règle, et nous contribuons ainsi à une ouverture par la construction de notre jeu de données. La procédure de collecte de données consiste en les points suivants, exécutés toutes les deux minutes par un script python :

- récupération de la page web brute donnant les informations de trafic
- parsing du code html afin de récupérer les identifiants des liens de trafic et les temps de parcours correspondants
- insertion des liens dans une base sqlite avec le temps courant.

Le script automatisé de collection des données continue d'enrichir la base au fur et à mesure du temps, permettant des développements futurs de ce travail sur un jeu de données plus large, et une réutilisation potentielle pour des travaux scientifiques ou opérationnels. La dernière version du jeu de données au format sqlite est disponible en ligne sous une Licence *Creative Commons*¹.

DESCRIPTION DES DONNÉES Une granularité de deux minutes a été obtenue pour une période de trois mois (de février 2016 à avril 2016 inclus). La granularité spatiale est en moyenne de 10km, les temps de trajet étant fournis pour les liens majeurs. Le jeu de données contient 101 liens. La variable brute utilisée est le temps de trajet effectif, à partir duquel il est possible de construire la vitesse de trajet et la vitesse relative de trajet, définie comme le rapport entre temps de trajet optimal (temps de trajet sans congestion, pris comme le temps minimal sur l'ensemble des pas de temps) et le temps de trajet effectif. La congestion est construite par inversion d'un fonction BPR simple avec exposant 1, i.e. en prenant $c_i = 1 - \frac{t_{i,\min}}{t_i}$ avec t_i temps de trajet effectif dans le lien i et $t_{i,\min}$ temps de trajet minimal.

¹ à l'adresse http://37.187.242.99/files/public/sytadin_latest.sqlite3

Méthodes and Résultats

VISUALISATION DES MOTIFS SPATIO-TEMPORELS DE CONGESTION

Notre approche étant entièrement empirique, une bonne connaissance des motifs existants pour les variables de traffic, en particulier de leur variations spatio-temporelles, est crucial pour guider toute analyse quantitative. En s'inspirant de la littérature étudiant la validation empirique de modèles, plus précisément les techniques de *Modélisation orientée-motifs* introduites par [GRIMM et al., 2005], nous nous intéressons au motifs macroscopiques à des échelles temporelles et spatiales données : d'une manière équivalente aux faits stylisés qui sont dans cette approches extraits d'un système avant de tenter de le modéliser, nous devons explorer les données de manière interactive dans le temps et l'espace afin d'identifier des motifs pertinents et les échelles associées. Une application web interactive a ainsi été implémentée pour explorer les données, à l'aide des packages R `shiny` et `leaflet`². Cela permet une visualisation dynamique des motifs de congestion sur l'ensemble du réseau ou dans une zone particulière grâce au zoom. L'application est accessible en ligne à l'adresse <http://shiny.parisgeo.cnrs.fr/transportation>. La Figure 12 présente une capture d'écran de l'interface. La conclusion majeure de l'exploration interactive des données est qu'une grande hétérogénéité spatiale et temporelle est la règle. Le motif temporel le plus récurrent, la périodicité journalière des heures de pointe, est perturbée pour une proportion non négligeable de jours. En première approximation, les heures creuses peuvent être approchées par une distribution localement stationnaire des flots, tandis que les heures de pointe sont trop courtes pour pouvoir impliquer la validation de l'hypothèse d'équilibre. Concernant l'espace, aucun motif spatial particulier n'émerge clairement. Cela signifie que dans le cas d'une validité de l'équilibre utilisateur statique, les méta-paramètres régissant son établissement doivent varier à des échelles temporelles plus courtes qu'un jour. Nous postulons au contraire que le système de traffic est loin de l'équilibre, en particulier pendant les heures de pointe pendant lesquelles des transitions de phase critiques à l'origine des embouteillages émergent.

VARIABILITÉ SPATIO-TEMPORELLE DES TRAJETS A la suite de l'exploration interactive des données, nous proposons de quantifier la variabilité spatiale des motifs de congestion pour valider ou invalider l'intuition que si l'équilibre existe par rapport au temps, il est fortement dépendant de l'espace et localisé. La variabilité spatio-temporelle des plus courts chemins de trajet est une première façon d'étudier la stationnarité des flots d'un point de vue de théorie des

² le code source de l'application et des analyses est disponible sur le dépôt ouvert du projet à
<https://github.com/JusteRaimbault/TransportationEquilibrium>

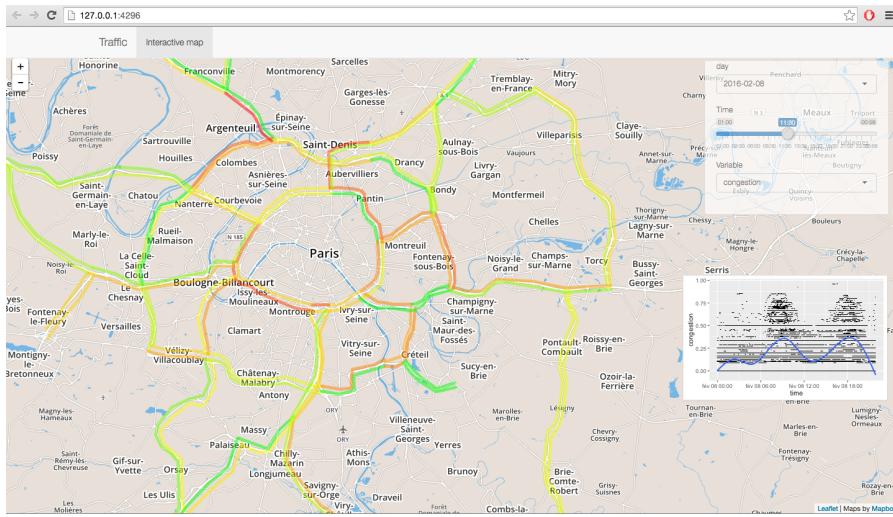


FIGURE 12 : Capture de l’application web permettant l’exploration spatio-temporelle des données de trafic pour la région Parisienne. Il est possible de choisir date et heure (précision de 15min sur un mois, réduite par rapport au jeu de données initial pour des raisons de performance). Un graphe résume les motifs de congestion pour la journée courante.

jeux. En effet, l’Equilibre Utilisateur Statique est la distribution stationnaire des flots sous laquelle aucun utilisateur ne peut augmenter son temps de trajet en changeant son itinéraire. Une forte variabilité spatiale des plus courts chemins sur de courtes échelles spatiales révèle ainsi une non-stationnarité, puisque un même utilisateur prendra un chemin complètement différent après un court laps de temps et ne contribuera plus au même flot que précédemment. Une telle variabilité est en effet observée sur un nombre non-négligeable de chemins pour chaque jour du jeu de données. La figure 13 montre un exemple de variation spatiale extrême d’un trajet pour une paire Origine-Destination particulière.

L’exploration systématique de la variabilité du temps de trajet sur l’ensemble du jeu de données, et des distances de trajet associées, confirme, comme présenté en figure , que la variation absolue du temps de trajet présente fréquemment une forte variation de son maximum sur l’ensemble des paires O-D, jusqu’à 25 minutes avec une moyenne temporelle locale autour de 10 minutes. La variabilité spatiale correspondante entraîne des détours allant jusqu’à 35km.

STABILITÉ DES MESURES DE RÉSEAU La variabilité des trajectoires potentielles observée dans la section précédente peu être confirmée par l’étude de la variabilité des propriétés du réseau. En particulier, les mesures topologiques de réseau capturent les motifs globaux dans

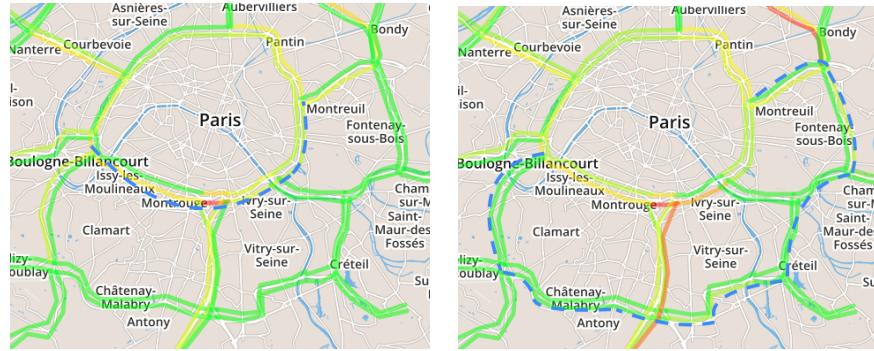


FIGURE 13 : Variabilité spatiale d'un plus court chemin en temps de trajet (trajet du plus court chemin en pointillé bleu). Dans un intervalle de seulement 10 minutes, entre le 11/02/2016 00 :06 (à gauche) et le 11/02/2016 00 :16 (à droite), le plus court chemin entre Porte d'Auteuil à l'ouest et Porte de Bagnolet à l'est, augmente en distance effective de $\simeq 37\text{km}$ (avec une augmentation du temps de trajet de seulement 6 minutes), à cause d'une forte perturbation sur le périphérique parisien.

un réseau de transport. Les mesures de centralité et de connectivité des noeuds sont des indicateurs classiques pour la description des réseaux de transport comme rappelé par [BAVOUX et al., 2005]. La littérature en transports a développé des mesures de réseau élaborées et opérationnelles, comme des mesures de robustesse pour identifier les liens critiques et mesurer la résilience globale du réseau aux perturbations (un exemple parmi d'autres est l'indice de *Robustesse du Réseau Effective* introduit dans [SULLIVAN et al., 2010]).

Plus précisément, nous étudions la centralité de chemin du réseau de transport, défini pour un noeud comme le nombre de plus courts chemins passant par celui-ci, i.e. par l'équation

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \rightarrow d)} \quad (3)$$

où V est l'ensemble des sommets du réseau de taille N , et $p(o \rightarrow d)$ est l'ensemble des noeuds sur le plus court chemin entre les sommets o et d (le plus court chemin étant calculé avec le temps de trajet effectif). Cette mesure de centralité est plus adaptée que d'autre dans notre cas, comme la centralité de proximité qui n'inclut pas la congestion potentielle comme la centralité de chemin.

Nous montrons en Figure 4 la variation relative absolue du maximum de la centralité de chemin, pour la même fenêtre temporelle

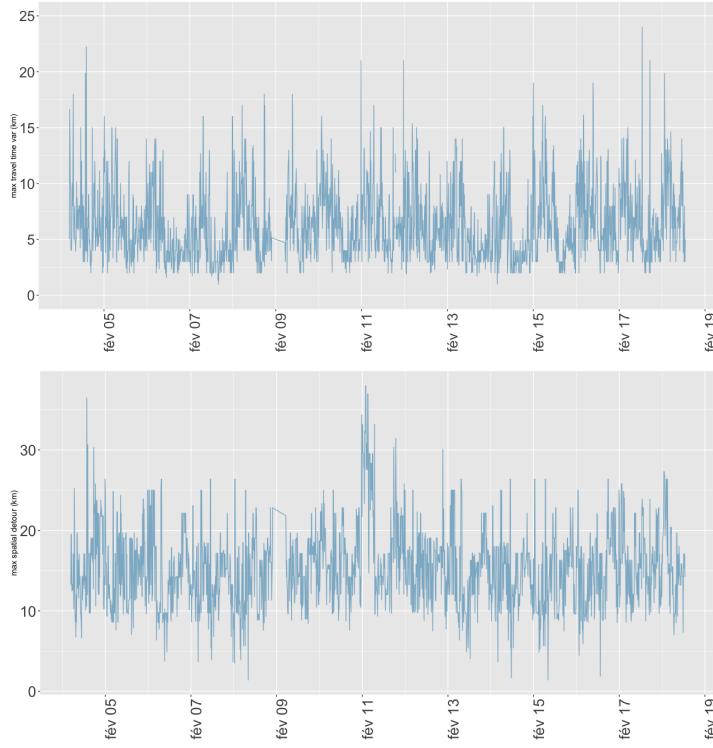


FIGURE 14 : Variabilité maximale du temps de trajet (en haut) en minutes et de la distance de trajet correspondante (en bas) pour un échantillon de deux semaines. Le graphe représente le maximum sur l'ensemble des paires Origine-Destination de la variabilité absolue entre deux pas de temps consécutifs. Les heures de pointe induisent une forte variabilité du temps de trajet, allant jusqu'à 25 minutes et une variabilité de distance jusqu'à 35km.

que les indicateurs empiriques précédents. Plus précisément, elle est définie par

$$\Delta b(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \quad (4)$$

où Δt est le pas de temps du jeu de données (la plus petite fenêtre temporelle sur laquelle une variabilité peut être capturée). Cette variation relative absolue a une signification directe : une variation de 20% (qui est atteinte un nombre significatif de fois comme montré en Figure 15) implique dans le cas d'une variation négative, qu'au moins cette proportion de trajectoires potentielles ont changé et que la potentielle congestion locale a décrue de la même proportion. Dans le cas d'une variation positive, un seul noeud a capturé au moins 20% des trajets. Sous l'hypothèse (qu'on ne tente pas de vérifier ici et qu'on peut également supposer non vérifiée comme montré par [ZHU et LEVINSON, 2010], mais que l'on utilise comme un outil pour donner une intuition sur la signification concrète de la variabilité de la centralité)

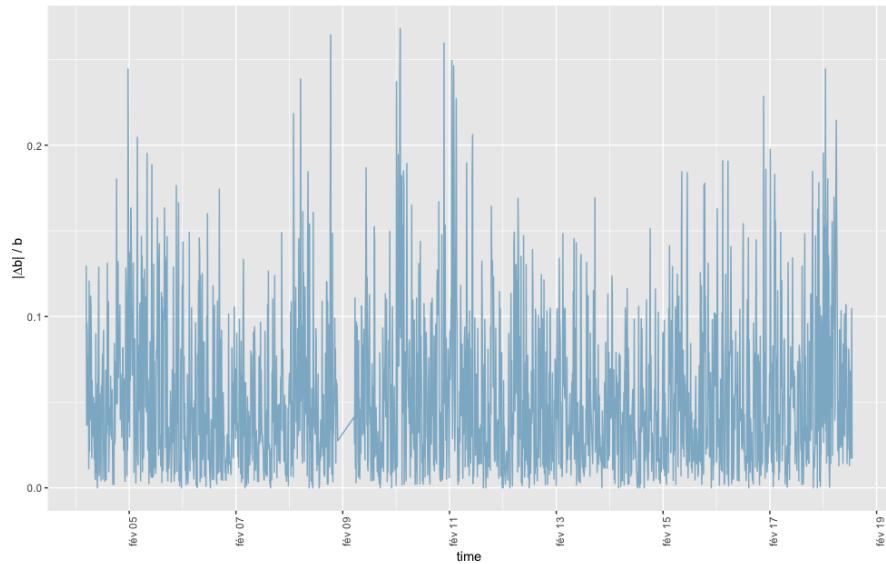


FIGURE 15 : Stabilité temporelle du maximum de la centralité de chemin. Le graphe montre dans le temps la dérivée normalisée du maximum de la centralité de chemin, qui capture ses variations relatives à chaque pas de temps. La valeur maximale de 25% correspond à de très fortes perturbations du réseau sur les liens correspondants, puisque cela implique qu'au moins cette proportion d'utilisateurs prenant le lien dans des conditions précédentes doivent prendre un trajet complètement différent.

que les utilisateurs choisissent rationnellement le plus court chemin, et supposant que la majorité des trajets est réalisées, une telle variation de la centralité implique une variation similaire dans les flots effectifs, conduisant à la conclusion qu'ils ne peuvent être stationnaires ni dans le temps (au moins sur une échelle plus grande que Δt) ni dans l'espace.

HÉTÉROGÉNÉITÉ SPATIALE DE L'ÉQUILIBRE Afin d'obtenir un point de vue différent sur la variabilité spatiale des motifs de congestion, nous proposons d'utiliser un indice d'auto-corrélation spatiale, l'indice de Moran (défini par exemple dans [TSAL, 2005]). Utilisé plus généralement en analyse spatiale, avec diverses applications allant de l'étude de la forme urbaine à la quantification de la ségrégation, il peut être appliqué à toute variable spatiale. Il permet d'établir des relations de voisinage et révèle la consistance spatiale locale d'un équilibre s'il est appliqué à une variable de trafic localisée. A un point donnée de l'espace, l'auto-corrélation locale pour la variable c est calculée par

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \quad (5)$$

où K est une constante de normalisation égale à la somme des poids spatiaux fois la variance de la variable et \bar{c} est la moyenne de la variable. Dans notre cas, nous choisissons des poids spatiaux de la forme $w_{ij} = \exp\left(\frac{-d_{ij}}{d_0}\right)$ avec d_0 distance typique de décroissance. L'auto-corrélation est calculée sur la congestion des liens, localisée au centre du lien. Elle capture ainsi les corrélations spatiales dans un rayon du même ordre que la distance de décroissance autour du point i . La moyenne sur l'ensemble des points fournit l'indice d'auto-corrélation spatiale I . Une stationnarité des flots devrait impliquer une stabilité temporelle de l'index.

La figure 16 présente l'évolution temporelle de l'auto-corrélation spatiale pour la congestion. Comme attendu, on observe une forte décroissance de l'auto-corrélation avec la distance de décroissance, à la fois sur l'amplitude et les moyennes temporelles. La forte variabilité temporelle implique de courtes échelles temporelles pour des fenêtres potentielles de stationnarité. Pour une distance de décroissance de 1km, en comparant l'auto-corrélation à la congestion (ajustée à l'échelle du graphe pour lisibilité), on observe que les fortes corrélations coïncident avec les heures creuses, tandis que les heures de pointe correspondent à une décroissance des corrélations. Notre interprétation, combinée avec la variabilité observée des motifs spatiaux, est que les heures de pointe correspondent à un comportement chaotique du système, puisque les bouchons peuvent émerger dans n'importe quel lien du réseau : la corrélation disparaît alors puisque l'espace des phases atteignables pour un système dynamique chaotique est rempli uniformément par les trajectoires, de façon équivalente à des vitesses relatives qui apparaîtraient comme aléatoires et indépendantes.

5.1.3 Discussion

Implications théoriques et pratiques des conclusions empiriques

Nous prétendons que les implications théoriques de ces résultats empiriques n'impliquent pas nécessairement un rejet total du cadre de l'Equilibre Utilisateur Statique, mais révèlent plutôt un besoin de plus fortes connexions entre la littérature théorique et les études empiriques. Si chaque nouveau cadre théorique introduit est généralement testé sur un cas ou plus, il n'existe pas de comparaisons systématiques de chacun sur des jeux de données de grande taille et variés, et pour des objectifs d'application différents (prédition du traffic, reproduction de faits stylisés, etc.), à l'image des revues systématiques qui sont la règle en évaluation thérapeutique par exemple. Cela implique cependant des pratiques de partage des données et des modèles plus larges que celles existant couramment. La connaissance précise des potentialités d'application d'un cadre donné peut induire des déve-

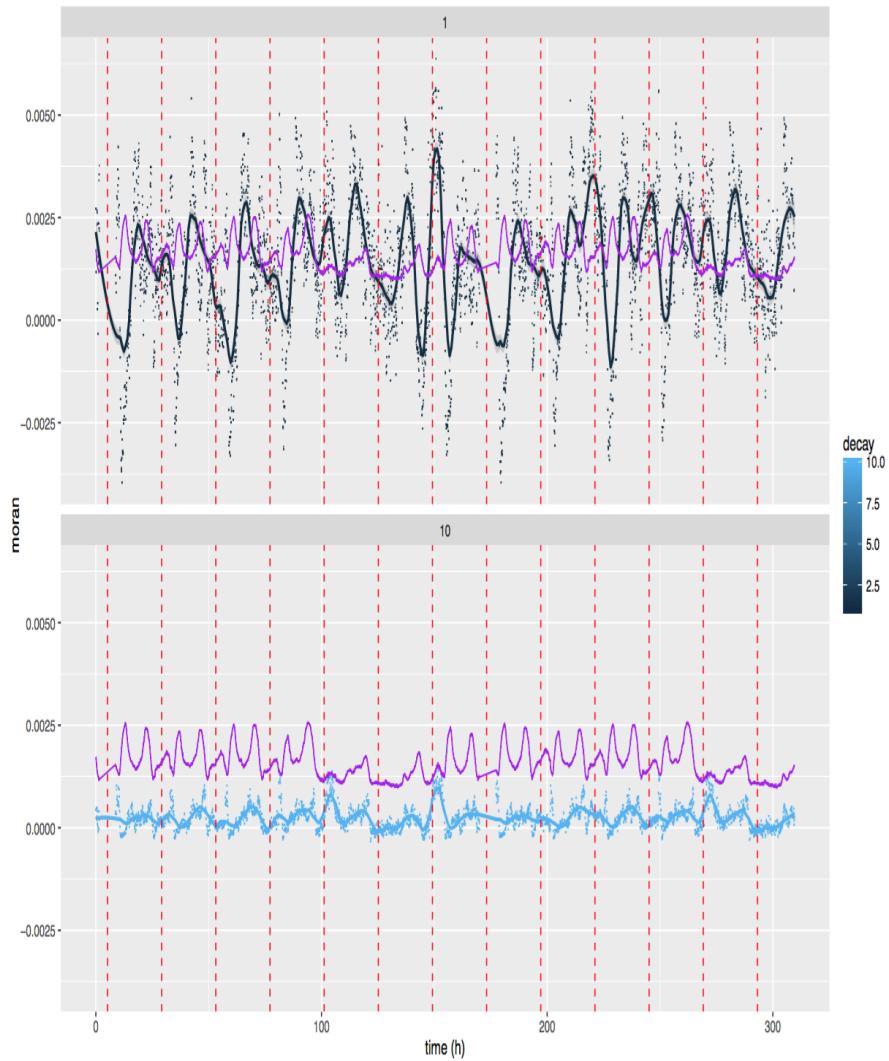


FIGURE 16 : Auto-corrélations spatiales pour les vitesses relatives sur deux semaines. Le graphe montre les valeurs de l'auto-corrélation dans le temps, pour des valeurs variables (1,10km) de la distance de décroissance. les valeurs intermédiaires de la distance de décroissance donnent une déformation relativement continue entre ces deux extrêmes. Les points sont lissés sur une fenêtre temporelle de 2h pour faciliter la lecture. Les lignes pointillées verticales correspondent à minuit de chaque jour. La courbe violette donne la vitesse relative, ajustée à l'échelle pour établir la correspondance entre les heures de pointe et les variations de l'auto-corrélation.

loppements inattendus comme l'intégration dans des modèles plus larges. L'exemple des études des interaction entre Transport et Usage du Sol (modèles *LUTI*) est une bonne illustration d'un cas où le EUS peut toujours être utilisé avec des motivations plus larges que la modélisation du traffic. [KRYVOBOKOV et al., 2013] décrit deux modèles *LUTI*, dont l'un inclut deux équilibres pour les modèles de transport à quatre temps et pour l'évolution de l'usage du sol (localisation des ménages et emplois), l'autre étant dynamique. La conclusion est que chaque modèle a ses avantages au regard de l'objectif poursuivi, et que le modèle statique peut être utilisé pour comparer des politiques sur le temps long, tandis que le modèle dynamique fournit de l'information plus précise à de plus petites échelles temporelles. Dans le premier cas, un module de transport plus compliqué aurait été plus difficile à inclure, ce qui est un avantage du EUS dans ce cas.

Concernant les applications pratiques, il semble naturel que les modèles statiques ne devraient pas être utilisés pour la prédiction et la gestion du traffic sur de petites échelles temporelles (semaine ou jour) et que des efforts doivent être faits pour implémenter des modèles plus réalistes. Cependant, l'utilisation des modèles par la communauté des ingénieurs et des planificateurs n'est pas directement reliée aux enjeux académiques et à l'état de l'art dans le domaine. Dans le cas particulier de la France et des modèles de mobilité, [COMMENGES, 2013a] a montré que les ingénieurs allaient jusqu'au point de construire des problèmes inexistant et d'implémenter les modèles correspondants qu'ils avaient importé d'un contexte géographique totalement différent (la planification aux Etats-Unis). L'utilisation d'un cadre ou d'un type de modèle a des raisons historiques qui peuvent être difficiles à surmonter.

Vers des interprétations de la non-stationnarité

Une hypothèse qu'on peut formuler concernant l'origine de la non-stationnarité des flots dans le réseau, au regard de l'exploration des données et des analyses quantitatives, est que le réseau est au moins la moitié du temps fortement congestionné et dans un état critique. Les heures creuses sont les plus grandes fenêtres temporelles potentielles de stationnarité spatiale et temporelle, mais couvre moins de la moitié du temps. Comme déjà interprété dans le comportement de l'indicateur d'auto-corrélation, un comportement chaotique pourrait être à l'origine d'une telle variabilité lors des heures congestionnées. A la manière d'un fluide supercritique qui condense sous une perturbation externe infinitésimale, l'état d'un lien peut qualitativement changer par un petit incident, produisant une perturbation du réseau qui se propage et peut même s'amplifier. L'effet direct des événements du traffic (incidents signalés ou accidents) ne peut pas être étudié sans source de données extérieure, et un enrichissement de la base de données dans cette direction pourrait être intéressante. Cela

permettrait d'établir la proportion de perturbations qui paraissent avoir un effet direct et quantifier un niveau de caractère critique de la congestion du réseau dans le temps, ou d'étudier plus précisément des phénomènes localisés comme les conséquences d'un incident de traffic sur la voie opposée.

Développements

Le travail futur pourra être planifié dans la direction d'une étude raffinée de la stabilité temporelle sur des zones du réseau, i.e. l'étude quantitative précise de la non-stationnarité des heures de pointes découverte ci-dessus. Pour cela nous proposons de calculer numériquement la stabilité de Liapounov du système dynamique régissant les flots de traffic, par l'intermédiaire d'algorithmes numériques comme ceux décrits par [GOLDHIRSCH, SULEM et ORSZAG, 1987]. La valeur des exposants de Liapounov fournit l'échelle de temps sur laquelle le système instable s'éloigne de l'équilibre. Leur comparaison avec la durée des heures de pointe et le temps de trajet moyen, sur différentes zones spatiales et différentes échelles, devrait fournir plus d'information sur une possible validité de l'hypothèse de stationnarité locale. Cette technique a déjà été introduite à une autre échelle dans les études de transport, comme e.g. [TORDEUX et LASSARRE, 2016] qui étudie la stabilité des modèles de régulation de vitesse à l'échelle microscopique pour éviter l'émergence de congestion.

D'autres directions de recherche peuvent consister en le test des autres hypothèses du EUS (comme le choix rationnel du plus court chemin, qui serait cependant difficile à tester à un tel niveau d'agrégation, impliquant l'utilisation de modèles de simulation calibrés et cross-validés sur le jeu de données pour comparer différentes hypothèses, sans toutefois nécessairement une validation ou invalidation directe de l'hypothèse), ou le calcul empirique des paramètres dans les cadres d'Equilibre Utilisateur Stochastique ou Dynamique.

Conclusion

Nous avons décrit une étude empirique ayant pour but une étude simple, mais selon notre point de vue nécessaire, de l'existence de l'équilibre utilisateur statique, plus précisément de sa stationnarité dans le temps et l'espace pour un réseau routier métropolitain principal. Un jeu de données de congestion du trafic est construite par collection de données, pour le réseau du Grand Paris sur 3 mois avec une granularité temporelle de 2 minutes. L'exploration interactive du jeu de données via une application web permettant la visualisation spatio-temporelle aide à guider les analyses quantitatives. La variabilité spatio-temporelle des plus courts chemins et de la topologie du réseau, en particulier la centralité de chemin, révèle que l'hypothèse de stationnarité ne tient généralement pas, ce qui est confirmé

par l'étude de l'auto-corrélation spatiale de la congestion du réseau. Nous suggérons que nos résultats soulignent un besoin général de plus grandes connexions entre les études théoriques et empiriques, puisque cette étude permet de chasser les incompréhensions théoriques sur l'Equilibre Utilisateur Statique, et guider le choix d'application potentielles.

★ ★

★

5.2 TRANSPORT ROUTIER ET DÉTERMINANTS DES COÛTS

La géographie des prix du carburant a de nombreuses applications variées, de son impact significatif sur l'accessibilité à son rôle comme indicateur d'équité territoriale et de politique de transports. Dans cette section, nous étudions les variations spatio-temporelles des prix du carburant aux Etats-Unis à une résolution très fine, par l'utilisation d'un nouveau jeu de données, donnant les prix journaliers sur deux mois pour une proportion significative des stations essence. Les données ont été collectées par l'intermédiaire d'une technologie de crawling à grande échelle élaborée spécifiquement, que l'on décrira. Nous étudions l'influence de variables socio-économiques, en utilisant des méthodes complémentaires : la Régression Géographique Pondérée pour tenir compte de la non-stationnarité spatiale, et une modélisation économétrique linéaire pour conditionner à l'Etat et tester des caractéristiques au niveau du Comté. La première fournit une portée spatiale optimale qui correspond globalement à l'échelle de stationnarité, et une influence significative des variables comme le revenu moyen ou le salaire par travail, avec un comportement spatial dont la non simplicité confirme l'importance des particularités géographiques. D'autre part, la modélisation multi-niveaux révèle un très fort effet Etat, alors que les caractéristiques spécifiques au Comté gardent un impact significatif. A travers la combinaison de ces méthodes, nous démontrons la superposition d'un processus de gouvernance avec un processus spatial socio-économique local. Nous discutons une application potentielle importante qui est l'élaboration de politiques de régulation automobiles localement paramétrées.

5.2.1 Contexte

Quels sont les déterminants des prix du carburant ? Par l'utilisation d'une nouvelle base de données des prix des carburants au niveau de la station, collectée pendant deux mois, nous explorons leur variabilité dans le temps et l'espace. Une variation du coût du carburant peut avoir de nombreuses causes, du prix brut du pétrole aux politiques fiscales locales et aux caractéristiques géographiques, chacun ayant des effets hétérogènes dans l'espace et le temps. Bien que l'évolution du prix moyen du carburant dans le temps soit un indicateur suivi avec attention et analysé par de nombreuses institutions financières, sa variabilité dans l'espace reste relativement non-explorée dans la littérature. Cependant, de telles différences peuvent refléter des variations dans des indicateurs socio-économiques plus indirects comme des inégalités territoriales, des singularités géographiques ou des préférences des consommateurs.

Il n'existe à notre connaissance pas de cartographie systématique dans le temps et l'espace des prix de vente à l'échelle d'un pays. La

raison principale est probablement que la disponibilité des données a pu être un obstacle important. Il est aussi probable que la nature de la question joue un rôle, puisque celle-ci se trouve à l'interface de plusieurs disciplines. Alors que les économistes étudient l'élasticité des prix et leur mesure dans différents marchés, la géographie des transports, par des méthodes comme les prix des transports intégrés aux modèles spatiaux, met une emphase plus grande sur la distribution spatiale que sur des mécanismes précis de marché. Toutefois, des exemples de travaux relativement liés peuvent être trouvés. Par exemple, [RIETVELD, BRUINSMA et VAN VUREN, 2001] étudie l'impact de différences de prix transfrontalières et leur implications pour une taxation spatiale graduelle aux Pays-Bas. A l'échelle du pays, [RIETVELD et WOUDENBERG, 2005] fournit des modèles statistiques pour expliquer les variabilité des prix entre les pays Européens. [MACHARIS et al., 2010] modélise l'impact d'une variation spatiale des prix sur les motifs d'intermodalité, ce qui implique que l'hétérogénéité spatiale des prix du carburant a un impact fort sur le comportement des utilisateurs. Avec une approche similaire par la géographie des transports, [GREGG et al., 2009] étudie la distribution spatiale des émissions à l'échelle des Etats américains. La géographie des prix du carburant a également d'importantes répercussions sur les coûts effectifs, comme le montre [COMBES et LAFOURCADE, 2005] en déterminant les coûts réels de transport pour les différentes aires urbaines françaises. De façon plus proche de notre travail, et en utilisant des données similaires en Accès Ouvert pour la France, [GAUTIER et SAOUT, 2015] étudie les dynamiques de transmission des prix bruts du pétrole aux prix de vente. Toutefois, ils n'introduisent pas de modèle spatial explicite de diffusion des prix et n'étudient pas de dynamiques spatio-temporelles.

Dans cette section nous adoptons une approche différente en procédant à une analyse spatiale exploratoire des prix du carburant aux Etats-Unis. Nous montrons que la majorité des variations s'observent entre les Comtés et non dans le temps, malgré les évolutions du baril brut pendant la période considérée. Nous employons pour cela une analyse spatiale de la distribution des prix. Les résultats majeurs obtenus sont les suivants : d'une part nous montrons l'existence de motifs spatiaux significatifs dans des grandes régions US, d'autre part nous montrons que même si la majorité des variations observées par les politiques des Etats, et en particulier le niveau de taxation, certaines caractéristiques à l'échelle du Comté restent significatives.

Dataset

Notre jeu de données contient l'information journalière des prix des carburants à l'échelle de la station essence pour l'ensemble du territoire US métropolitain. Ces informations sont construites à partir des prix reportés par les utilisateurs et couvre pratiquement l'ensemble

des station essence aux Etats-Unis. Nous commençons par décrire la collection des données et donnons des statistiques de ce jeu de données nouveau.

Collection de données hétérogènes à grande échelle

La disponibilité de nouveaux types de données a conduit à des évolutions significatives dans de nombreuses disciplines (e.g. l'analyse des réseaux sociaux en ligne ([TAN et al., 2013])) à la géographie (e.g. les nouvelles approches de la mobilité urbaine ou les perspectives de ville plus "intelligentes" ([BATTY, 2013a])) en incluant l'économie pour laquelle la disponibilité de données exhaustives à l'échelle individuelle ou de l'entreprise est vu comme une révolution dans le champ. La plupart des études impliquant ces nouvelles données sont à l'interface des disciplines concernées, ce qui est à la fois un avantage mais aussi une source de complications. Par exemple les malentendus entre physique et sciences urbaines décrites par [DUPUY et BENGUIGUI, 2015] sont en particulier causées par des attitudes différentes au regard des données non conventionnelles ou des interprétations et ontologies différentes pour celles-ci. La collection et l'utilisation des nouvelles données est donc devenu un enjeu essentiel en sciences sociales. La construction des tels jeux de données est cependant loin d'être évidente de par la nature incomplète et bruitée de la donnée. Des outils techniques spécifiques doivent être implantés mais sont souvent conçus pour surmonter un problème donné et sont difficiles à généraliser. Nous développons un tel outil qui remplit les contraintes suivantes typiques de la collection de données à grande échelle : (i) un niveau raisonnable de flexibilité et de généralité; (ii) une performance optimisée par la collection parallélisée; (iii) l'anonymat des jobs de collection pour éviter le plus possible tout biais dans le comportement de la source de données. L'architecture, à un assez haut niveau, a la structure suivante :

- Un ensemble indépendant des tâches fait tourner en continu des proxies socks pour envoyer les requêtes via tor.
- Un manager suit les tâches de collection en cours, réparti la collection entre les sous-tâches et en lance des nouvelles lorsque cela s'avère nécessaire.
- Les sous-tâches peuvent être toute application prenant comme argument les adresses de destination, elles procèdent à la collecte, au parsage et au stockage des données collectées.

L'application est ouverte et ses modules sont réutilisables : le code source est disponible sur le dépôt du projet.³ Nous avons construit notre jeu de données en utilisant l'outil en continu pendant deux

³ à <https://github.com/JusteRaimbault/EnergyPrice>

TABLE 2 : Statistiques descriptives des prix des carburants (\$ par gallon)

Moyenne	Dev. Std.	p10	p25	p50	p75	p90
2.28	0.27	2.02	2.09	2.21	2.39	2.65

mois pour collecter des données crowdsourcées disponibles de diverses sources en ligne.

Jeu de données

Le jeu de données contient autour de $41 \cdot 10^6$ observations uniques des prix de vente au niveau de la station, s'étendant sur une période du 10 janvier 2017 au 19 mars 2017, correspondant à 118,573 station service uniques. Pour chacune, nous disposons d'une localisation géographique précise (résolution à la ville). En moyenne nous avons 377 informations de prix par station. Les prix correspondent à un mode d'achat unique (par carte de crédit, les autres modes comme l'argent liquide représentant moins de 10% sur des jeux tests, ils ont été abandonnés dans le jeu de données final) et quatre types de carburant possibles : Diesel (18% des observations), Regular (34%), Midgrade (24%) et Premium (24%). La meilleure couverture des stations est pour le carburant Regular avec en moyenne 4,629 données de prix par Conté. Nous choisissons pour cette raison de concentrer l'étude sur ce type de carburant, en gardant à l'esprit que des développements futurs avec le jeu de données pourraient inclure des analyses comparatives des types de carburant. Notre jeu de données final contient ainsi 14,192,352 observations provenant de 117,155 stations service, suivies pendant 68 jours. Nous agrégeons de plus les données par jour, en prenant la moyenne du prix observé par gallon, pour obtenir un panel de 5,204,398 observations station-jour.⁴ La table 2 donne des statistiques descriptives basiques sur les données de prix, montrant que la distribution des prix est fortement concentrée avec une faible skewness (le ratio du 99th au 1st quantiles est 1.6). Enfin, dans l'analyse spatiale, nous utiliserons également des données socio-économiques au niveau du Conté, disponible par le US Census Bureau. Nous utiliserons les plus récentes disponibles (ce qui dans la plupart des cas implique d'utiliser le Census de 2010).

⁴ Le panel n'est pas équilibré puisque les pris ne sont pas reportés chaque jour pour chaque station. Une station moyenne possède l'information de prix pour 44 jours (sur 68).

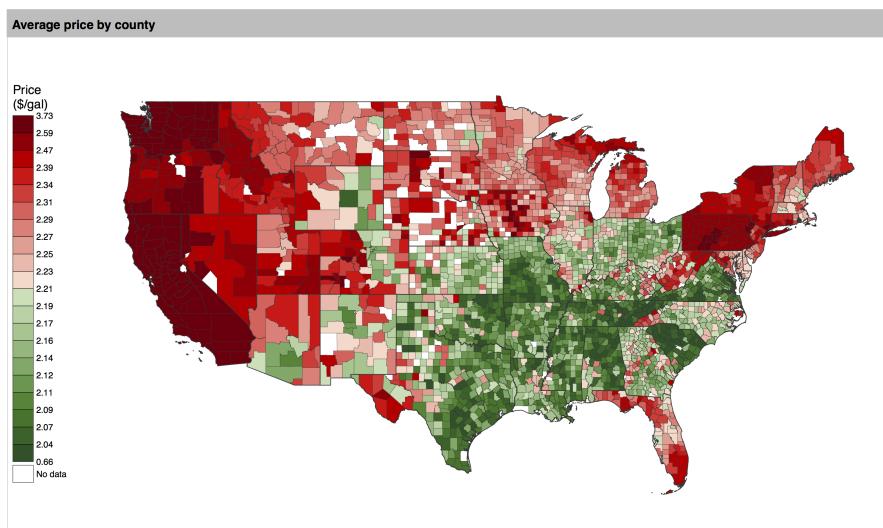


FIGURE 17 : Carte du prix moyen par Comté, carburant régulier, moyenne prise sur l'ensemble de la période.

5.2.2 Résultats

Motifs spatio-temporels des prix

Avant de se consacrer à une étude plus systématique de la variation des prix des carburants, nous proposons une première introduction exploratoire pour donner une idée de sa structure spatio-temporelle. Cette exercice est une étape cruciale pour guider les analyses suivantes, mais aussi pour comprendre leurs implications dans le contexte géographique. Afin d'explorer les données, nous construisons une application web basique permettant de cartographier les données dans l'espace et le temps. Elle est disponible à . Nous montrons également de carte au niveau du Comté à la figure 17 pour le prix moyen sur l'ensemble de la période. On voit clairement apparaître des motifs régionaux, avec les régions du centre sud et du sud est ayant les prix les plus bas et la côte Pacifique et le nord est les prix les plus hauts. Bien évidemment, une carte agrégée sur l'ensemble de la période n'apporte guère d'information sur les variations temporelles des données. Comme nous allons le montrer plus en détails par la suite, la majorité des variations des prix des carburants a lieu dans l'espace. une décomposition de la variance des prix donne seulement 11% de la variance totale expliquée par les variations intra-station. De la même manière, le coefficient de corrélation de rang de Spearman entre le prix des stations pour le carburant regular entre le premier jour du jeu de données et le dernier jour est de 0.867, et l'hypothèse nulle que ces deux informations sont indépendantes est fortement rejetée.

Puisque la majorité de la variation des prix est inter-station, nous nous intéressons maintenant principalement aux corrélations spatiales. Nous conduisons l'analyse à l'échelle du Conté pour diverse raisons. D'une part une décomposition des prix des carburants inter et intra-Conté montre que plus de 85% de la variance est inter-Conté, d'autre part car la localisation des stations n'est pas assez fiable pour permettre une granularité plus fine, et enfin car la majorité des variables socio-économiques est à ce niveau. Nous étudions donc l'autocorrelation spatiale des prix à l'échelle du Conté. L'autocorrelation spatiale peut être vue comme une indicateur d'hétérogénéité spatiale que nous mesurons par l'index de Moran ([TSAI, 2005]), avec des poids spatiaux de la forme $\exp(-d_{ij}/d_0)$ avec d_{ij} étant la distance entre les entités spatiales i et j , et d_0 un paramètre de décroissance donnant la portée spatiale des interactions que l'estimation prend en compte. Nous montrons en Fig. 18 ses variation pour chaque jour ainsi que comme fonction du paramètre de décroissance. Les fluctuations dans le temps de l'index de Moran journalier pour les valeurs basses et moyennes du paramètre de decay, confirme les spécificité géographiques au sens de régimes de corrélation changeant localement. Celles-ci sont logiquement atténuées pour les longues portées, puisque les correlations des prix diminuent avec la distance. Le comportement de l'autocorrelation spatiale en fonction du paramètre de decay est particulièrement intéressant : nous observons une premier changement de régime autour de 10km (d'un régime constant à un régime linéaire par morceau), et une seconde transition importante autour de 1000km, les deux consistants sur des fenêtres temporelles à la semaine. Nous postulons que celles-ci correspondent au échelles spatiales typiques des phénomènes observés : le régime bas serait les spécificités locales et l'intermédiaire le processus au niveau de l'Etat. Ce comportement confirme que les pris sont non-stationnaires dans l'espace, et que pour cette raison des techniques statistiques appropriées doivent être utilisées pour étudier les variables jouant un rôle à différents niveaux. Les deux parties suivantes suivent cette idée et étudient des variables explicatives potentielles des prix locaux du carburant, utilisant deux techniques différentes qui correspondent à deux paradigmes complémentaires : la régression géographique pondérée qui met l'emphase sur les effets de voisinage, et des régressions multi-niveaux prenant en compte les limites administratives.

Régression Géographique Pondérée

La question de la non-stationnarité des processus géographiques a toujours été une source d'analyses agrégées biaisées ou de mauvaises interprétations lorsque des conclusions générales sont appliquées à des cas locaux. Pour le prendre en compte dans les modèles statistiques, de nombreuses techniques ont été proposées, parmi lesquelles la simple mais très élégante Régression Géographique Pon-

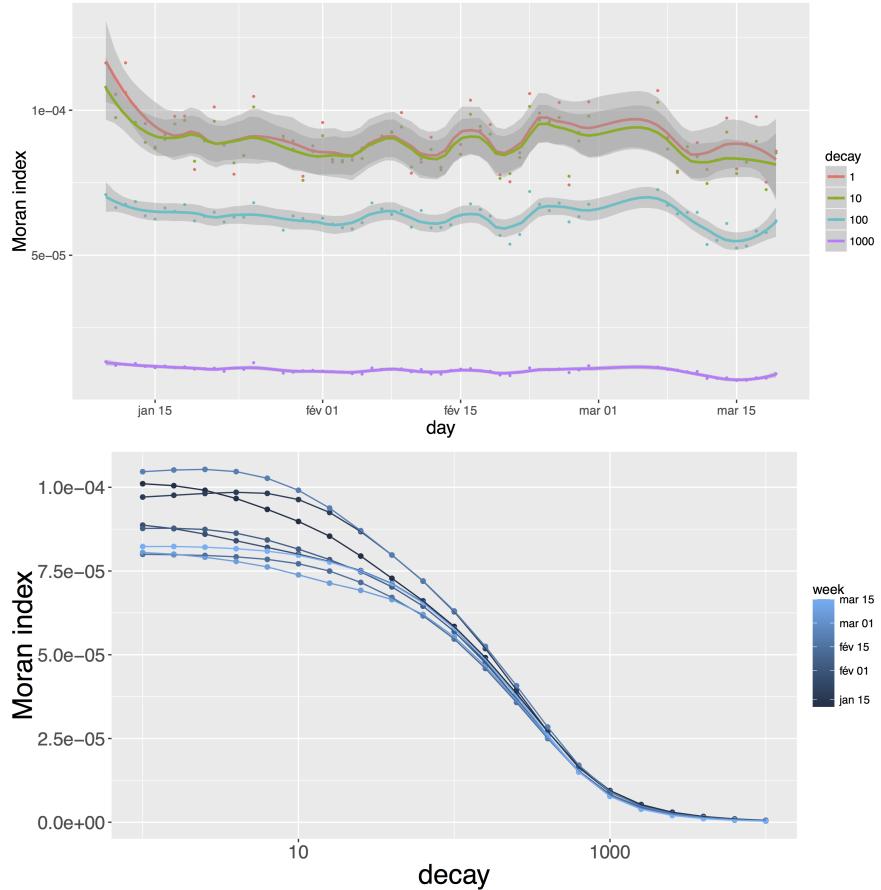


FIGURE 18 : Comportement de l'index d'autocorrelation spatiale de Moran.
 (Gauche) Evolution dans le temps de l'index de Moran, calculé sur des fenêtres journalières, pour différentes valeurs du paramètre de décroissance. (Droite) Index de Moran en fonction du paramètre de décroissance, calculé sur des fenêtres hebdomadaires.

dérée (GWR), qui estime des régressions non-stationnaires en pondérant les observations dans l'espace de manière similaire aux techniques d'estimation de densité par noyaux. Elle a été introduite dans un article séminal par [BRUNSDON, FOTHERINGHAM et CHARLTON, 1996] et a été utilisée et développée en conséquence depuis. L'avantage considérable de cette technique est qu'une portée spatiale optimale au sens de la performance du modèle peut être déduite pour dériver un modèle qui traduit des effets des variables variant dans l'espace, révélant ainsi des effets locaux qui peuvent se produire à différentes échelles spatiales ou à travers les frontières. Nous procédons à un multi-modeling pour trouver le meilleur modèle et le noyau ainsi que la portée spatiale associés. Plus précisément, nous suivons les étapes suivantes : (i) tous les modèles linéaires potentiels à partir des cinq variables candidates sont générés (revenu, population, salaire par emploi, emploi par tête, emplois); (ii) pour chaque modèle et chaque forme de noyau candidate (exponentiel, gaussien, bisquare, escalier), nous déterminons la portée optimale au sens à la fois de la cross-validation et du critère d'Information d'Akaike corrigé (AICc) qui quantifie l'information contenue dans le modèle; (iii) nous ajustons les modèles avec cette portée. Nous choisissons le modèle avec le meilleur AICc, en l'occurrence $\text{price} = \beta \cdot (\text{income}, \text{wage}, \text{percapjobs})$ pour une portée de 22 voisins et un noyau Gaussien,⁵ avec un AICc de 2,900. La différence médiane d'AICc avec l'ensemble des autres modèles est 122. Le coefficient de détermination global est 0.27, ce qui est relativement bon en comparaison du meilleur R-squared de 0.29 (obtenu pour le modèle avec l'ensemble des variables, qui surfe clairement avec un AICc de 3010; de plus la dimension effective est inférieure à 5 puisque 90% de la variance est expliquée par les trois premières composantes principales pour les variables normalisées).

Les coefficients et le R-squared local pour le meilleur modèle sont montrés en Fig. 19. La distribution spatiale des résidus (qui n'est pas montrée ici), semble globalement distribuée aléatoirement, ce qui confirme d'une certaine façon la cohérence de l'approche. En effet, si une structure géographique distinguable était trouvée dans les résidus, cela signifierait que le modèle géographique ou les variables considérées ont échoués à traduire la structure spatiale. Nous pouvons à présent proposer une interprétation des structures spatiales obtenues. Tout d'abord, la distribution spatiale de la performance du modèle révèle des régions où ces indicateurs socio-économiques simples expliquent relativement bien les prix, et celles-ci sont localisées sur la côte ouest, la frontière sud, la région nord-est des lacs à la côte est, et une bande de Chicago au sud du Texas. Les coefficients correspondants ont des comportements différents selon les

⁵ on note que la forme du noyau n'a pas plus d'influence tant que des fonctions décroissantes graduellement sont utilisées.

zones, suggérant différents régimes.⁶ Par exemple, l'influence du revenu dans chaque région semble s'inverser quand la distance à la côte augmente (du nord au sud-est dans l'ouest, du sud au nord au Texas, de l'est à l'ouest & l'est), ce qui pourrait témoigner de différentes spécialisations économiques. Au contraire, le changement de régime pour les salaires montre une rupture notable entre l'ouest (sauf autour de Seattle) et le centre et l'est, qui ne correspond pas directement à des politiques d'Etat locales puisque le Texas est coupé en deux par exemple. De la même façon, les emplois par capita montrent une opposition entre est et ouest, qui pourrait être due par exemple à des différences culturelles. Ces résultats sont toutefois difficiles à interpréter directement, et doivent être compris comme la confirmation que les particularités géographiques importent, puisque les régions diffèrent dans le régime du rôle de chacune des variables socio-économiques simples. Une connaissance plus précise pourrait être obtenue par des études géographiques ciblées incluant des études de terrain qualitatives et des analyses quantitatives, qui sont au delà de la portée de cette étude exploratoire et laissée à une éventuelle recherche future.

Enfin, nous extrayons l'échelle spatiale des processus étudiés, c'est à dire en calculant la distribution de la distance aux plus proches voisins avec la portée optimale. On obtient approximativement une distribution log-normale, de médiane 77km et d'interquartile 30km. Nous interprétons cette échelle comme l'échelle de stationnarité spatiale du processus de prix en relation avec les agents économiques, qui peut également être comprise comme la portée des marchés cohérents de compétition entre les stations service.

Régressions multi-niveaux

Comme notre base initiale permet de regarder au niveau des variables $x_{i,s,c,t}$, le prix du carburant au jour t, dans la station i, dans l'Etat s et dans le Comté c, nous commençons par estimer des régressions à effets fixes en grande dimension, suivant le modèle :

$$x_{i,s,c,t} = \beta_s + \varepsilon_{i,s,c,t} \quad (6)$$

$$x_{i,s,c,t} = \beta_c + \varepsilon_{i,s,c,t} \quad (7)$$

$$x_{i,s,c,t} = \beta_i + \varepsilon_{i,s,c,t} \quad (8)$$

$$(9)$$

Où $\varepsilon_{i,s,c,t}$ contient une erreur idiosyncratique et un effet fixe jour. Cette première analyse confirme que la majorité de la variance peut être expliquée par un effet fixe Etat et que d'intégrer des niveaux plus fins a un effet négligeable sur la performance du modèle mesurée par le R-squared.

⁶ Nous commentons leur comportement dans les zones où le modèle a une performance minimale, que nous fixons arbitrairement à un R-squared local de 0.5.

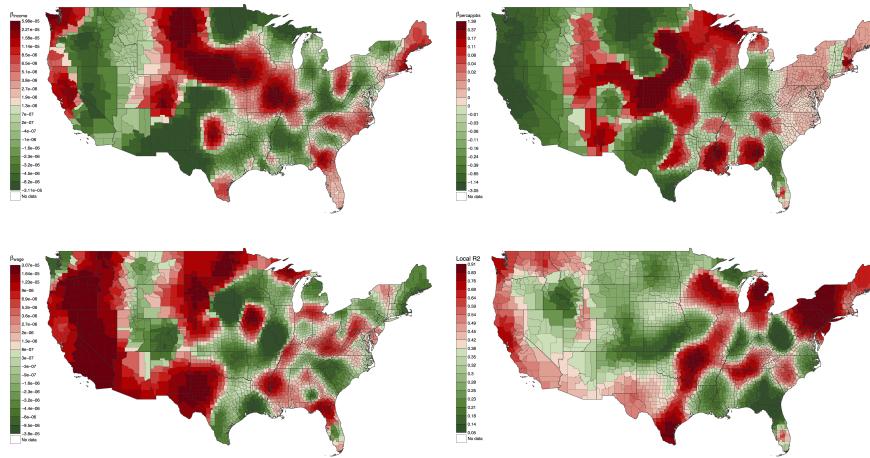


FIGURE 19 : Résultats des analyses GWR. Pour le meilleur modèle au sens de l'AICc, les cartes donnent la distribution spatiale des coefficients estimés, dans l'ordre de gauche à droite et de haut en bas, β_{income} , $\beta_{percapjobs}$, β_{wage} , et finalement les valeurs du R-squared local.

Nous nous tournons à présent vers une analyse différente, visant à capturer les variables explicatives qui rendent compte des variations spatiales du carburant. Nous considérons le modèle linéaire suivant :

$$\log(x_i) = \beta_0 + X_i \beta_1 + \beta_{s(i)} + \varepsilon_i, \quad (10)$$

où x_i dénote le prix moyen mesuré du carburant dans le Conté i agrégé sur l'ensemble des jours, X_i est un ensemble de variables spécifiques au Conté et $s(i)$ est l'état dans lequel se trouve le Conté de telle façon que $\beta_{s(i)}$ capture toute la variation spécifique aux Etats. Enfin ε_i est un terme d'erreur satisfaisant $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ si $s(i) \neq s(j)$. ce regroupement de l'erreur standard au niveau de l'état est motivé par les résultats de la partie précédente, montrant que l'autocorrélation spatiale des prix du carburant au niveau de l'état est toujours potentiellement forte. Cette spécification vise à capturer les effets de variables socio-économiques variées au niveau du Conté après que l'effet fixe Etat aie été retiré. Les résultats sont présentés en Table 3. La première colonne montre que la regression du logarithme des prix sur un effet fixe Etat est déjà suffisant pour expliquer 74% de la variance. Cela est majoritairement du aux taxes sur les carburants qui sont fixées au niveau de l'Etat aux Etats-Unis. En fait, une régression du log-prix sur le niveau de taxe donne un R-squared de 0.33%. Les variables explicatives restantes montrent que les Contés urbains denses ont des prix plus élevés, mais que le prix décroît avec la population. Ce résultat paraît raisonnable, les zones désertiques ayant en moyenne des prix plus hauts. Les prix augmentent avec le revenu

total, décroissent avec le niveau de pauvreté et décroisse avec le niveau de vote pour un candidat républicain. Ce dernier point suggère un lien circulaire : les Contés qui utilisent beaucoup la voiture auront tendance à voter pour un politicien qui promouvra des politiques favorable à son usage. L'ajout de ces variables explicatives augmente légèrement le R-squared, ce qui suggère que même après avoir enlevé l'effet fixe Etat, la prix du carburant peut être expliqué par des caractéristiques socio-économiques locales.

5.2.3 *Discussion*

SUR LA COMPLÉMENTARITÉ DES MÉTHODES ÉCONOMÉTRIQUES ET DES MÉTHODES D'ANALYSE SPATIALE Un aspect important de cette contribution est méthodologique. Nous montrons que pour explorer un nouveau panel de données, les géographes et les économistes prennent des approches différentes, menant à des conclusions génériques similaires par des chemins différents. Des études ont déjà combiné les GWR et les régressions multi-niveau ([CHEN et TRUONG, 2012]), ou les ont comparées en terme de performance de modèle ou de robustesse ([LEE, KANG et KIM, 2009]). Nous prenons ici un point de vue multi-disciplinaire et combinons des approches répondant à des questions différentes, GWR ayant pour but de trouver des variables explicatives précises et de mesurer le rôle de l'auto-corrélation spatiale, tandis que les modèles économétriques expliquent plus précisément les effets des différents facteurs à plusieurs niveaux (Etat, Conté) mais prennent ces caractéristiques géographiques comme exogènes. Nous postulons que les deux sont nécessaires pour comprendre toutes les dimensions du phénomène étudié.

PROPOSITION DE POLITIQUES DE RÉGULATION LOCALISÉES Une autre application de ce type d'analyse est d'aider à une meilleure conception de politiques de régulation de la voiture. Les problèmes environnementaux et de santé requièrent de nos jours un usage raisonnable de celle-ci, dans les villes avec le problème de la pollution atmosphérique, mais aussi globalement pour réduire les émissions de CO₂. [FULLERTON et WEST, 2002] montre qu'une taxation des carburants et des voitures peut être équivalente à une taxation des émissions. [BRAND, ANABLE et TRAN, 2013] souligne le rôle des incitations pour une transition vers des transports décarbonés. Cependant, de telles mesures ne peuvent pas être uniformes d'un Etat à l'autre ou même entre les Contés, pour des raisons évidentes d'équité territoriale : des zones avec des caractéristiques socio-économiques différentes ou avec différentes aménités doivent contribuer selon leur possibilité et préférences. La connaissance des dynamiques locales des prix et leur déterminants, ce en quoi notre étude est une étape préliminaire, peut

TABLE 3 : Régressions au niveau du Conté

	(1)	(2)	(3)	(4)	(5)
Density	0.016*** (0.002)	0.016*** (0.001)	0.016*** (0.001)	0.015*** (0.001)	
Population (log)	-0.007*** (0.001)	-0.040*** (0.011)	-0.041*** (0.011)	-0.039*** (0.010)	
Total Income (log)		0.031*** (0.010)	0.031*** (0.010)	0.027*** (0.009)	
Unemployment		0.001 (0.001)	0.000 (0.001)	0.000 (0.001)	
Poverty		-0.028** (0.011)	-0.030*** (0.011)	-0.029** (0.011)	
Percentage Black			0.000*** (0.000)	-0.000 (0.000)	
Vote GOP				-0.072*** (0.015)	
R-squared	0.743	0.767	0.774	0.776	0.781
N	3,066	3,011	3,011	3,011	3,011

Notes : Cette table donne les résultats d'une régression des Moindres Carrés Ordinaire pour le modèle présenté en équation (10). La densité est mesurée comme le nombre d'habitants au mile-carré et le revenu total est donné en dollars. La pauvreté est mesurée comme le nombre de personnes sous le seuil de pauvreté par habitants. On étudie aussi l'influence du pourcentage de personnes noires et de la part de personnes ayant voté pour Donald Trump aux élections de 2016. La régression inclut un effet fixe Etat. Les erreurs standard robustes, agrégées au niveau de l'état, sont données entre parenthèses. ***, ** and * indiquent respectivement les niveaux de significativité 0.01, 0.05 and 0.1.

être une voie vers des régulations localisées prenant en compte la configuration socio-économique et inclure un critère d'équité.

Conclusion

Nous avons décrit une première étude exploratoire des prix des carburants aux US dans le temps et l'espace, utilisant une nouvelle base de données au niveau de la station s'étendant sur deux mois. Notre premier résultat est de montrer la grande hétérogénéité spatiale des processus de prix, par une exploration interactive des données et des analyses d'auto-corrélation. Nous procémons à deux études complémentaires des déterminants potentiels : GWR révèle des structures spatiales et des particularités géographiques, and fournit une échelle caractéristique des processus autour de 75km ; les régressions multi-niveaux montrent que même si la majorité des variations sont expliquées par les caractéristiques des Etats, et majoritairement par le niveau de taxation fixé par l'Etat, il existe toujours des spécificités socio-économiques au niveau du Comté qui peuvent expliquer la variation spatiale des prix du carburant.

* * *

*

5.3 TRANSACTIONS IMMOBILIÈRES ET GRAND PARIS

5.3.1 Contexte

Des aspects très variés des territoires sont concernés par l’interaction avec les réseaux. Dans nos études précédentes, les aspects économiques et financiers du foncier et l’immobilier n’ont pas été considérés. Il s’agit cependant d’éléments cruciaux des dynamiques territoriales et sont étudiés de manière intensive dans des champs comme l’analyse territoriale ou l’économie urbaine : par exemple, [HOMOCIANU, 2009] étudie les choix résidentiels des ménages pour comprendre les interactions entre usage du sol et transport. Nous proposons ici d’utiliser entre autres une base de données de transactions immobilières pour la région parisienne sur les 20 dernières années, avec une granularité temporelle de 2 ans et coordonnées spatiales exactes. [GUÉROIS et LE GOIX, 2009] l’utilise par exemple pour établir une typologie des dynamiques spatiales du marché immobilier parisien.

Notre approche peut être comprise comme une recherche de signes précurseurs de rupture de potentiels du réseau : en effet, si des dynamiques territoriales intrinsèques anticipent l’arrivée d’une nouvelle station de transports en commun, les implications seront bien différentes du cas où celle-ci conduit ces variables après sa construction. L’interprétation en termes “d’effets structurants” sera notamment très différente. Nous appliquons ici la méthode de causalités spatio-temporelles

La région métropolitaine de Paris est en train de connaître de grandes mutations, avec la mise en place d’une gouvernance métropolitaine et de nouvelles infrastructures de transport par exemple. La construction d’un réseau de métro en rocade permettant des liaisons de banlieue à banlieue est un besoin ancien, et a mené à plusieurs propositions sur lesquelles se sont opposés l’Etat et la Région au tournant des années 2010 [DESJARDINS, 2010]. Le projet Arc Express [STIF, 2010], porté par la Région et plus axé sur une égalité des territoires, contrastait avec les propositions initiales de Réseau du Grand Paris visant à relier des “clusters d’excellence” en dépit d’un possible effet tunnel. La solution finalement adoptée (voir le dernier schéma directeur [SDRIF, 2013]) est un compromis et permet un rééquilibrage est-ouest de l’accessibilité [BEAUCIRE et DREVELLE, 2013]. Nous proposons d’étudier les relations entre différentiel d’accessibilité pour chaque projet, et variables liées au foncier (transactions immobilières) et socio-économiques. En effet, les liens entre nouvelles lignes et évolution du foncier sont parfois remarquables [DAMM et al., 1980].

5.3.2 Cas d'étude

Données

Les données des transactions immobilières sont fournies par la base BIENS (Chambre des Notaires d'Ile de France, base propriétaire). Le nombre de transactions utilisables après nettoyage est de 862360, se répartissant sur l'ensemble des IRIS, pour une plage temporelle couvrant de 2003 à 2012 incluses. Les données par IRIS pour population et revenu (revenu médian et indice de Gini) proviennent de l'INSEE. Les données de réseau ont été vectorialisées à partir des cartes des projets (voir Fig. 20 pour les projets). Les temps de trajets sont calculés par transport en commun uniquement, avec des valeurs standard pour les vitesses moyennes des différents modes (RER 60km.h⁻¹, Transilien 100km.h⁻¹, Metro 30km.h⁻¹, Tramway 20km.h⁻¹). La matrice des temps est calculée depuis l'ensemble des centroïdes des IRIS vers l'ensemble des centroïdes des communes. Ceux-ci sont reliés au réseau par des connecteurs à la gare la plus proche, de vitesse 50km.h⁻¹ (trajet en voiture). Les analyses sont implémentées intégralement en langage R [R CORE TEAM, 2015a] et l'ensemble des données, du code source et des résultats sont disponibles sur un dépôt git ouvert⁷.

Résultats

Nous calculons pour chaque projet, le différentiel ΔT_i d'accessibilité en temps moyen de trajet à partir de chaque IRIS en comparaison à celui dans le réseau sans le projet, défini par $T_i = \sum_k \exp -t_{ik}/t_0$ avec k communes, t_{ik} temps de trajet, et t_0 paramètre d'atténuation. A chaque projet est associée une date⁸, correspondant environ à l'année d'annonce mature du projet, restant toutefois arbitraire car difficile d'une part à déterminer précisément, un projet n'émergeant pas d'un coup du jour au lendemain, et d'autre part pouvant correspondre à des réalités différentes d'apprentissage du projet par les différents agents économiques (nous faisons donc l'hypothèse réductrice mais nécessaire d'une diffusion sur la majorité des agents dans un temps inférieur à l'année). Nous étudions les corrélations décalées de cette variable avec les variations ΔY_{ij} des variables socio-économiques suivantes : population, revenu médian, indice de Gini des revenus, prix moyen des transactions immobilières et montant moyen des crédits immobiliers. Un test de Fisher est effectué pour chaque estimation, et la valeur est fixée nulle si celui-ci n'est pas significatif ($p < 0.05$ de

⁷ A l'adresse

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/SpatioTempCausality/GrandParis>
Les données de la base BIENS ne sont fournies que de manière agrégée à l'IRIS et pour les variables de prix et de crédit, pour des raisons de fermeture contractuelle de la base brute.

⁸ 2006 pour Arc Express, 2008 pour le Réseau du Grand Paris, 2010 pour le Grand Paris Express

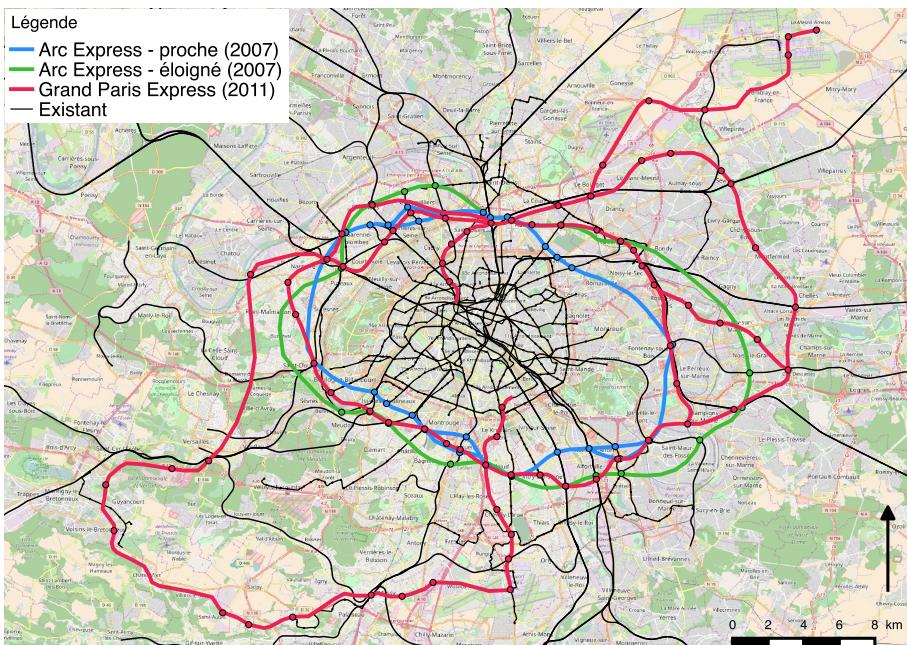


FIGURE 20 : Projets de transport successifs de la métropole du Grand Paris. Nous montrons les deux alternatives du projet Arc Express porté par la région, et le Grand Paris Express (GPE) porté par l'état. Le Réseau du Grand Paris, précurseur du GPE, n'est pas montré ici pour des raisons de visibilité à cause de sa proximité avec celui-ci.

manière classique). L'étude avec accessibilité généralisées au sens de Hansen a également été menée mais moins intéressante car très peu sensible à la composante mobilité (réseau et atténuation) par rapport aux variables elle-même, informe uniquement sur des relations entre celles-ci et n'est donc pas présentée ici. Nous présentons en Fig. 21 les résultats pour l'ensemble des réseaux et variables. Il est remarquable tout d'abord de noter l'existence d'effets significatifs pour l'ensemble des variables. Des valeurs plus basses du paramètre t_0 donnent des corrélations plus fortes en valeur absolue, révélant une possible plus grande importance de l'accessibilité locale sur les dynamiques territoriales. Le comportement de la population montre un pic très détaché correspondant à 2008, laissant supposer un impact du plus vieux projet d'Arc Express sur la croissance de la population, l'effet des autres projets serait alors fallacieux de par leur proximité dans les grands tronçons : cela impliquerait que les zones où ils diffèrent fondamentalement comme le Plateau de Saclay ne soient que très peu sensibles au projet de transport, ce qui confirmerait l'aspect artificiel planifié du développement de ce territoire. Concernant les revenus, on observe un comportement similaire mais négatif, ce qui impliquerait un appauvrissement lié à l'augmentation de l'accessibilité, mais qui semble toutefois s'accompagner d'une baisse des inégalités. Enfin, comme

attendu les prix immobiliers sont tirés par l'arrivée potentielle des nouveaux réseaux, effet qui disparaît à deux ans pour le Grand Paris Express, suggérant une bulle immobilière passagère. Nous démontrons ainsi l'existence de liens de correlations retardées complexes qu'on nomme causalités en ce sens, entre dynamiques territoriales et dynamiques anticipées des réseaux. Une compréhension plus fine des processus à l'oeuvre est au delà de la portée de cet article, car supposerait des études de terrain qualitatives, des études de cas ciblées, etc. Cet exemple illustre cependant le caractère opérationnel de notre méthode sur un cas d'étude réel.

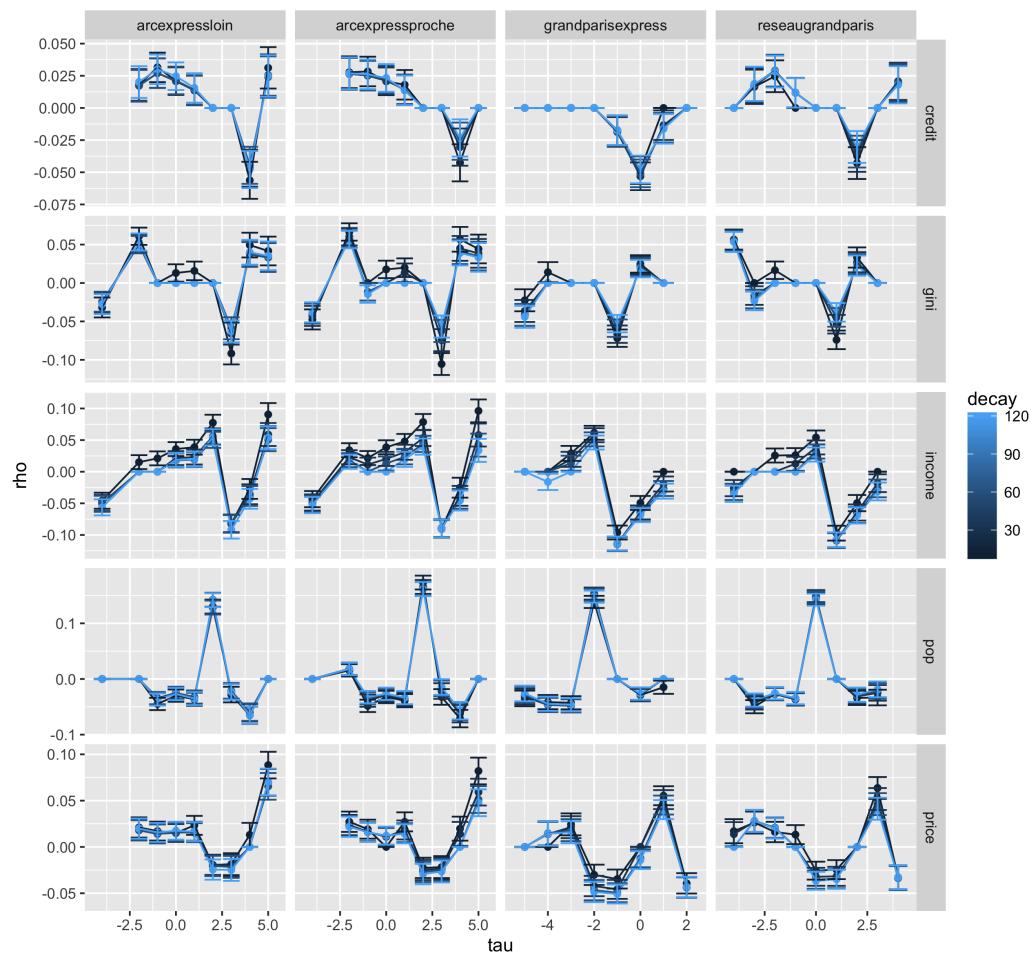


FIGURE 21 : Corrélations retardées empiriques. Les graphiques donnent la valeur de la corrélation entre le différentiel d'accessibilité en temps de trajet moyen ΔT pour chaque projet (en colonnes) et le différentiel des différentes variables socio-économiques et de transactions immobilières (en lignes), pour différentes valeurs du paramètre d'atténuation (decay). Les barres d'erreur donnent l'intervalle de confiance à 95%.

5.3.3 Discussion

Diffusion spatio-temporelle

L'application de notre approche doit être menée précautionneusement concernant le choix des échelles, processus et objets d'étude. Typiquement, elle ne sera pas du tout adaptée à la quantification de processus spatio-temporels dont l'échelle temporelle de diffusion est de l'ordre de celle de la fenêtre d'estimation : l'hypothèse de stationnarité est basique. On peut proposer de procéder à des estimations par fenêtres glissantes, mais il faudrait ensuite élaborer une technique de correspondance spatiale pour traquer la propagation des phénomènes. Un exemple d'application concrète à l'impact thématique fort serait une caractérisation d'une composante fondamentale de la Théorie Evolutive des Villes, la diffusion hiérarchique de l'innovation entre les villes [PUMAIN, 2010], en analysant les potentielles dynamiques spatio-temporelles des classifications de brevets comme celle introduite par [BERGEAUD, POTIRON et RAIMBAULT, 2017]. Il faut noter toutefois qu'il s'agit de questions méthodologiques relativement ouvertes, dont une des manifestations est le lien potentiel entre le caractère non-ergodique des systèmes urbains [PUMAIN, 2012b] et une caractérisation ondulatoire de ces processus.

Regression Géographique Pondérée

Une autre direction de développement et d'applications potentiels se révèle en se tournant vers l'échelle plus locale, et d'explorer une hybridation avec les techniques de Regression Géographique Pondérée [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998]. La détermination par validation croisée ou Critère d'Akaike d'une portée spatiale optimale pour la performance de ce type de modèles pourrait être adaptée dans notre cas pour déterminer une échelle locale optimale sur laquelle les corrélations retardées sont les plus significatives, ce qui permettrait de s'extraire du problème de la non-stationnarité prioritairement par l'aspect spatial.

* * *

*

CONCLUSION DU CHAPITRE

Cette collection d'études empiriques nous permet à la fois d'illustrer par des cas concrets nos considérations générales sur les réseaux et territoires, mais aussi de clarifier les échelles et ontologies qu'il nous est pertinent d'utiliser. Comme développé par 5.1, l'échelle microscopique dans le temps et l'espace, pour les objets du traffic routier ici, présente des dynamiques chaotiques, rendant peu réaliste l'intégration de cette échelle dans des modèles qui rendraient comptes d'interactions à de plus grandes échelles. Si cet aspect est pris en compte, c'est généralement sous la forme de congestion, qui est agrégée à une échelle supérieure et pour laquelle soit les conséquences des propriétés chaotiques ont été lissées (ce qui peut être un problème pour les modèles d'équilibre), soit elles sont calibrées empiriquement et l'échelle inférieure n'a donc pas d'ontologie dans le modèle. Nous prendrons ce parti dans nos modèles impliquant un transport routier. Ensuite dans 5.2, toujours concernant le réseau de transport routier, mais selon le point de vue d'un ancrage nodal dans les territoires par les stations essence, en relation avec diverses caractéristiques socio-économiques de ces territoires, nous démontrons d'une part l'existence d'échelles endogènes, correspondant à l'échelle mesoscopique et l'échelle macroscopique, et d'autre part la complexité des processus d'interaction mis en jeu de par leur non-stationnarité déjà démontrée en 4.1 mais aussi par la superposition d'effets territoriaux locaux à des effets liés à la gouvernance. La dernière section 5.3 permet de conforter ces conclusions de par l'existence d'effet causaux significatif à une échelle mesoscopique dans le temps et dans l'espace. Nous ferons ainsi les choix de modélisation de séparer les échelles, les modèles macroscopiques (comme celui déjà introduit en 4.3) visant à capturer la non-stationnarité en regardant la dynamique à un niveau supérieur en étudiant des variables simples, les modèles mesoscopiques visant à traduire les processus de morphogenèse locaux. Ceux-ci seront introduits dans le chapitre suivant. L'existence d'effets causaux nous confortent dans la recherche de régimes de causalité dans les modèles de co-évolution, comme introduits en 4.2, ce qui sera fait en chapitre 8. Enfin, les processus de gouvernance feront l'objet d'une attention particulière dans la modélisation proposée en 8.3.

* * *

*

6

MORPHOGENÈSE URBAINE

6.1 UNE APPROCHE INTERDISCIPLINAIRE DE LA MORPHOGENÈSE

Une première étape essentielle est la clarification de ce qui est entendu par le terme de morphogenèse. Initialement introduit en biologie, son transfert à d'autres champs s'est accompagné d'une déformation des concepts associés. Nous adaptons et traduisons ici le texte de [ANTELOPE et al., 2016] qui propose une entrée interdisciplinaire sur la morphogenèse. Brique essentielle de nos constructions, il est en effet crucial de lui donner une armature rigoureuse et claire. Nous prenons le parti d'une vision croisée, dans l'idée d'un perspectivisme appliqué comme introduit en section ??, pour obtenir des concepts aussi génériques et larges que possible.

OBJECTIF

CONTEXTE

6.1.1 Revues

Biologie du Développement

Intelligence Artificielle

Sciences Territoriales

Sciences Sociales et Psychologie

Autres

UNE APPROCHE MATHÉMATIQUE René Thom a développé dans *Stabilité Structurelle et Morphogenèse* [THOM, 1974] une théorie de la dynamique des systèmes, la théorie des catastrophes, qui étudie en profondeur l'impact de la structure topologique des variétés de l'espace des phases sur les dynamiques du système. Soit M une variété différentiable, dans laquelle l'état du système (m, \dot{m}) est embarqué. On suppose l'existence d'un ensemble fermé K appelé *Ensemble de Catastrophe*. Le type topologique de K est en fait déterminé de manière endogène par la dynamique du système (dans les cas simples, il réfère au types "classiques" d'attracteurs/points fixes que l'on connaît usuellement : points et cycles limites). Quand m traverse K , le système rencontre un changement *qualitatif* dans sa forme, ce qui constitue la base de la *morphogenèse*. Cette théorie abstraite de la morphogenèse est indépendante de la nature du système étudié, sa contribution principale étant de classifier les catastrophes locales qui occurrent lors de la morphogenèse. La différentiation et la richesse des motifs ont ainsi une explication géométrique à travers les types topologiques des catastrophes.

AUTOPOIÈSE ET MORPHOGENÈSE La notion d'*autopoïèse* exprime la capacité d'un système à s'auto-reproduire. Une caractérisation basique est une frontière semi-perméable produite par le système et la capacité à reproduire ses composants. Une définition plus générale est proposée par BOURGINE et STEWART dans [BOURGINE et STEWART, 2004] :

6.1.2 *Synthèse*

6.1.3 *Discussion*

6.2 MORPHOGENÈSE URBAINE PAR AGRÉGATION-DIFFUSION

C : (Florent) avant de dire ce que tu fais en couplant cela, on a besoin de connaitre un panorama des différentes approches pour modéliser les dynamiques urbaines

C : (Florent) il y a d'autres modèles couplant diffusion et croissance comme Fischer-Skellam (?) [BOSCH, METZ et DIEKMANN, 1990]

C : (Florent) effective dimension of urban system : sens ?

C : (Florent) n_d est ce un paramètre du modèle ?

C : (Florent) on indicator choice : pourquoi ce choix ? qu'en attends tu ?

C : (Florent) on scala implementation : si la question computationnelle prend de l'importance dans la thèse, il faudra donner de la matière

C : (Florent) on LHS : qu'est ce que cela veut dire : force brute ?

C : (Florent) figure/par on real data : ordre ? real data : lesquelles ? qu'est ce que ça veut dire ? tout le monde n'a pas lu Cottineau :)

C : (Florent) on Moran vs Entropy : pouvait-on prévoir zone impossible ? (dispersion faible incompatible avec forte autocorrelation spatiale

C : (Florent) on PCA objective : pourquoi se fixer cet objectif si particulier ?

C : (Florent) on calibration process : là encore, pourquoi ce choix ? tout est discutable : il faut expliciter

C : (Florent) on multi-scale dvlpmnt : pas clair ce que tu as en tête ici : je ne sais pas si tu auras le temps de creuser cela, mais pour du multi-scalaire, les schémas sont très aidant car c'est vite difficile à visualiser

6.2.1 Contexte

6.2.2 Modèle et Résultats

6.2.3 Discussion

6.3 GÉNÉRATION DE CONFIGURATIONS TERRITORIALES CORRÉLÉES

Cette section vise à explorer un couplage séquentiel (ou couplage simple) du modèle de génération de densité précédent avec une heuristique de croissance de réseau. Nous explorons par là un espace faisable de corrélations entre les mesures de réseau et les mesures morphologiques.

6.3.1 *Données Géographiques corrélées de Densité et de Réseau*

Contexte

L'une des inspirations et applications de la présente démarche est la génération de données synthétiques, par exemple pour alimenter les analyses de sensibilité à la configuration spatiale présentées en section 3.2. En géographie, l'utilisation de données synthétiques est plus généralement axée vers l'utilisation de population synthétiques au sein de modèles basés agents, comme par exemple des modèles de mobilité, des modèles *LUTI* [PRITCHARD et MILLER, 2009]. On peut également citer des méthodes d'analyse spatiales qui s'en rapprochent : par exemple, l'extrapolation d'un champ spatial continu à partir d'un échantillon discret, par une estimation par noyaux par exemple, peut être compris comme la génération d'un jeu de données synthétiques (même si ce n'est pas le point de vue initial, comme pour la Regression Géographique Pondérée [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998], dans laquelle les noyaux de taille variables n'interpolent pas des données au sens propre mais extrapolent des variables abstraites représentant l'interaction entre variables explicites). Dans le domaine de la modélisation en géographie quantitative, dans le cas de *modèles jouets* ou de modèles hybrides, une configuration initiale cohérente est souvent essentielle : un ensemble de configurations initiales possibles est alors un jeu de données synthétiques sur lesquelles le modèle est testé : le premier modèle Simpop [SANDERS et al., 1997], pionnier d'une famille de modèles par la suite paramétrisés par des données réelles, pourrait rentrer dans ce cadre mais était lancé sur une spatialisation synthétique unique. De même, il a été souligné la difficulté de générer une configuration initiale pour une infrastructure de transport dans le cas du modèle SimpopNet [SCHMITT, 2014], alors qu'il s'agit un point essentiel dans la connaissance du comportement du modèle. Il a récemment été proposé de contrôler systématiquement les effets de la configuration spatiale sur le comportement de modèles de simulation spatialisés [COTTINEAU et al., 2015b], méthodologie pouvant être interprétée comme un contrôle par données statistiques spatiales. L'enjeu est de pouvoir alors distinguer effets propres dus à la dynamique intrinsèque du modèle, d'effet particuliers dus à la structure géographique du cas d'application. Celui-ci

est crucial pour la validation des conclusions issues des pratiques de modélisation et simulation en géographie quantitative.

Formalisation

Dans notre cas, nous proposons de générer des systèmes de villes représentés par une densité spatiale de population $d(\vec{x})$ et la donnée d'un réseau de transport $n(\vec{x})$, représenté de façon simplifiée, pour lesquels on serait capable de contrôler les corrélations entre mesures morphologiques de la densité urbaine et caractéristiques du réseau. La question de l'interaction entre territoire et réseaux de transport est un sujet d'étude classique [OFFNER et PUMAIN, 1996], mais toujours majoritairement ouvert, extrêmement complexe et difficile à quantifier [OFFNER, 1993]. Une modélisation dynamique des processus impliqués devrait apporter des connaissances sur ces interactions ([BRETAGNOLLE, 2009a], p. 162-163). Dans ce cadre, nous développons un couplage *simple* (c'est à dire sans boucle de rétroaction) entre un modèle de morphogenèse urbaine et un modèle de génération de réseau.

MODÈLE DE DENSITÉ Les modèle de densité est celui décrit et exploré dans la section précédente. Nous l'utilisons pour la génération conditionnelle du réseau.

MODÈLE DE RÉSEAU D'autre part, on est capable de générer par un modèle N un réseau de transport planaire à une échelle équivalente, étant donné une distribution de densité. La génération du réseau étant conditionnée à la donnée de la densité, les estimateurs des indicateurs de réseau seront conditionnels d'une part, et d'autre part les formes urbaines et du réseau devraient nécessairement être corrélées, les processus n'étant pas indépendants. La nature et la modularité de ces corrélations selon la variation des paramètres des modèles restent à déterminer par l'exploration du modèle couplé.

La procédure de génération heuristique de réseau est la suivante :

1. Un nombre fixé N_c de centres qui seront les premiers noeuds du réseau est distribué selon la distribution de densité, suivant une loi similaire à celle d'agrégation, i.e. la probabilité d'être distribué sur une case est $\frac{(P_i/P)^\alpha}{\sum(P_i/P)^\alpha}$. La population est ensuite répartie selon les zones de Voronoi des centres, un centre cumulant la population des cases dans son emprise.
2. Les centres sont connectés de façon déterministe par percolation entre plus proches clusters : tant que le réseau n'est pas connexe, les deux composantes connexes les plus proches au sens de la distance minimale entre chacun de leurs sommets sont connectées par le lien réalisant cette distance. On obtient alors un réseau arborescent.

3. Le réseau est alors modulé par ruptures de potentiels afin de se rapprocher de formes réelles. Plus précisément, un potentiel d'interaction gravitaire généralisé entre deux centres i et j est défini par

$$V_{ij}(d) = \left[(1 - k_h) + k_h \cdot \left(\frac{P_i P_j}{P^2} \right)^\gamma \right] \cdot \exp \left(-\frac{d}{r_g(1 + d/d_0)} \right)$$

où d peut être la distance euclidienne $d_{ij} = d(i, j)$ ou la distance par le réseau $d_N(i, j)$, $k_h \in [0, 1]$ un poids permettant de changer le rôle des population dans le potentiel, γ régissant la forme de la hiérarchie selon les valeurs des populations, r_g distance caractéristique de décroissance et d_0 paramètre de forme. Cette forme de potentiel suppose d'une part que l'atténuation de l'interaction due à la distance est indépendante de la force de l'interaction due aux poids (hypothèse standard des modèles gravitaires) ; d'autre part qu'un terme constant du à la distance peut prendre plus ou moins de poids (pondération par k_h) ; et enfin que la fonction de distance prend comme paramètre une distance caractéristique, mais aussi un paramètre de forme, permettant par exemple de contrôler la décroissance sur les faibles distances.

4. Un nombre $K \cdot N_L$ de nouveaux liens potentiels est pris comme les couples ayant le plus grand potentiel pour la distance euclidienne ($K = 5$ est fixé).
5. Parmi les liens potentiels, N_L sont effectivement réalisés, qui sont ceux ayant le plus faible rapport $V_{ij}(d_N)/V_{ij}(d_{ij})$: à cette étape seul l'écart entre distance euclidienne et distance par le réseau compte, ce rapport ne dépendant plus des populations et étant croissant en d_N à d_{ij} fixé.
6. Le réseau est planarisé par création de noeuds aux intersections éventuelles créées par les nouveaux liens.

Notons que la construction du modèle de génération est heuristique, et que d'autres types de modèles comme un réseau biologique auto-généré [TERO et al., 2010], une génération par optimisation locale de contraintes géométriques [BARTHÉLEMY et FLAMMINI, 2008] ou un modèle de percolation plus complexe que celui utilisé, peuvent le remplacer, et permettraient la création de boucles dans le réseau. Ainsi, dans le cadre d'une architecture modulaire où le choix entre différentes implémentations d'une brique fonctionnelle peut être vue comme méta-paramètre [COTTINEAU, CHAPRON et REUILLOU, 2015], on pourrait choisir la fonction de génération adaptée à un besoin donné (par exemple proximité à des données réelles, contraintes sur les relations entre indicateurs de sortie, variété de formes générées, etc.).

ESPACE DES PARAMÈTRES L'espace des paramètres du modèle couplé¹ est constitué des paramètres de génération de densité $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$ (voir section 6.2 ; on s'intéresse pour simplifier au rapport entre population et taux de croissance, i.e. le nombre d'étapes nécessaires pour générer, et on fixe la population totale) et des paramètres de génération de réseau $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. On notera $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

INDICATEURS On quantifie la forme urbaine et la forme du réseau, dans le but de moduler la corrélation entre ces indicateurs. La forme est définie par un vecteur $\vec{M} = (r, \bar{d}, \varepsilon, a)$ donnant auto-corrélation spatiale (indice de Moran), distance moyenne, entropie, hiérarchie (voir [LE NÉCHET, 2015] pour une définition précise de ces indicateurs). Les mesures de la forme du réseau $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ sont, avec le réseau noté (V, E) ,

- Centralité moyenne \bar{c} , définie comme la moyenne de la *betweenness-centrality* (normalisée dans $[0, 1]$) sur l'ensemble des liens.
- Longueur moyenne des chemins \bar{l} définie par $\frac{1}{d_m} \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} d_N(i, j)$ avec d_m distance de normalisation prise ici comme la diagonale du monde $d_m = \sqrt{2}N$. **C : (Florent) pas la même que tantôt ? que souhaitez tu capturer ? pour le dire autrement qu'est ce qui fera que tu trouveras ton modèle "bon" ou pas ?**
- Vitesse moyenne [BANOS et GENRE-GRANDPIERRE, 2012], qui correspond à la performance du réseau par rapport au trajet à vol d'oiseau, définie par $\bar{s} = \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} \frac{d_{ij}}{d_N(i, j)}$.
- Diamètre du réseau $\delta = \max_{ij} d_N(i, j)$

COVARIANCE ET CORRELATION On s'intéressera à la matrice de covariance croisée $\text{Cov}[\vec{M}, \vec{G}]$ entre densité et réseau, estimée sur un jeu de n réalisations à paramètres fixés $(\vec{M}[D(\vec{\alpha})], \vec{G}[N(\vec{\alpha})])_{1 \leq i \leq n}$ par l'estimateur standard non-biaisé. On prend comme correlation associée la correlation de Pearson estimée de la même façon.

Implémentation

Le couplage des modèles génératifs est effectué à la fois au niveau formel et au niveau opérationnel, c'est à dire qu'on fait interagir des implémentations indépendantes. Pour cela, le logiciel OpenMole [REUILLOU,

¹ Le couplage faible permet de limiter le nombre total de paramètres puisqu'un couplage fort incluant des boucles de retroaction comprendrait nécessairement des paramètres supplémentaires pour régler la forme et l'intensité de celles-ci. Pour espérer le diminuer, il faudrait concevoir un modèle intégré, ce qui est différent d'un couplage fort dans le sens où il n'est pas possible de figer l'un des sous-systèmes pour obtenir un modèle de l'autre correspondant au modèle non-couplé.

LECLAIRE et REY-COYREHOURCQ, 2013] utilisé pour l'exploration intensive, offre le cadre idéal de par son langage modulaire permettant de construire des *workflows* par composition de tâches à loisir et de les brancher sur divers plans d'expérience et sorties. Pour des raisons opérationnelles, le modèle de densité est implémenté en langage `scala` comme un plugin d'OpenMole, tandis que la génération de réseau est implémentée en langage basé-agent NetLogo [WILENSKY, 1999], ce qui facilite l'exploration interactive et construction heuristique interactive. Le code source est disponible pour reproductibilité sur le dépôt du projet².

Résultats

L'étude du modèle de densité seul est développée dans la section précédente. Il est notamment calibré sur les données de la grille européenne de densité, sur des zones de 50km de côté et de résolution 500m pour lesquelles les valeurs réelles des indicateurs ont été calculées pour l'ensemble de l'Europe. D'autre part, une exploration brutale du modèle permet d'estimer l'ensemble des sorties possibles dans des bornes raisonnables pour les paramètres (grossièrement $\alpha \in [0.5, 2]$, $N_G \in [500, 3000]$, $P_m \in [10^4, 10^5]$, $\beta \in [0, 0.2]$, $n_d \in \{1, \dots, 4\}$). La réduction à un plan de l'espace des objectif par une Analyse en Composantes Principales (variance expliquée à deux composantes $\simeq 80\%$ **C : (Florent) est-ce beaucoup ? est-ce bien que ce soit beaucoup ?**) permet d'isoler un nuage de points de sorties recouvrant assez fidèlement le nuage des points réels, ce qui veut dire que le modèle est capable de reproduire morphologiquement l'ensemble des configurations existantes.

A densité donnée, l'exploration de l'espace des paramètres du modèle de réseau suggèrent une assez bonne flexibilité sur des indicateurs globaux \tilde{G} , ainsi que de bonnes propriétés de convergence. Pour une étude du comportement précis, voir l'appendice donnant les regressions traduisant le comportement du modèle couplé. Dans le but d'illustrer la méthode de génération de données synthétiques, l'exploration a été orientée vers l'étude des correlations.

Etant donné la grande dimension relative de l'espace des paramètres, une exploration par grille exhaustive est impossible. On utilise un plan d'expérience par criblage (hypercube latin), avec les bornes indiquées ci-dessus pour $\tilde{\alpha}_D$ et pour $\tilde{\alpha}_N$, on a $N_C \in [50, 120]$, $r_g \in [1, 100]$, $d_0 \in [0.1, 10]$, $k_h \in [0, 1]$, $\gamma \in [0.1, 4]$, $N_L \in [4, 20]$. Concernant le nombre de réplications du modèle pour chaque valeur des paramètres, moins de 50 sont nécessaires pour obtenir sur les indicateurs des intervalles de confiance à 95% de taille inférieure aux déviations standard. Pour les correlations, une centaine donne des IC (obtenus par méthode de Fisher) de taille moyenne 0.4, on fixe donc $n = 80$

² à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic>

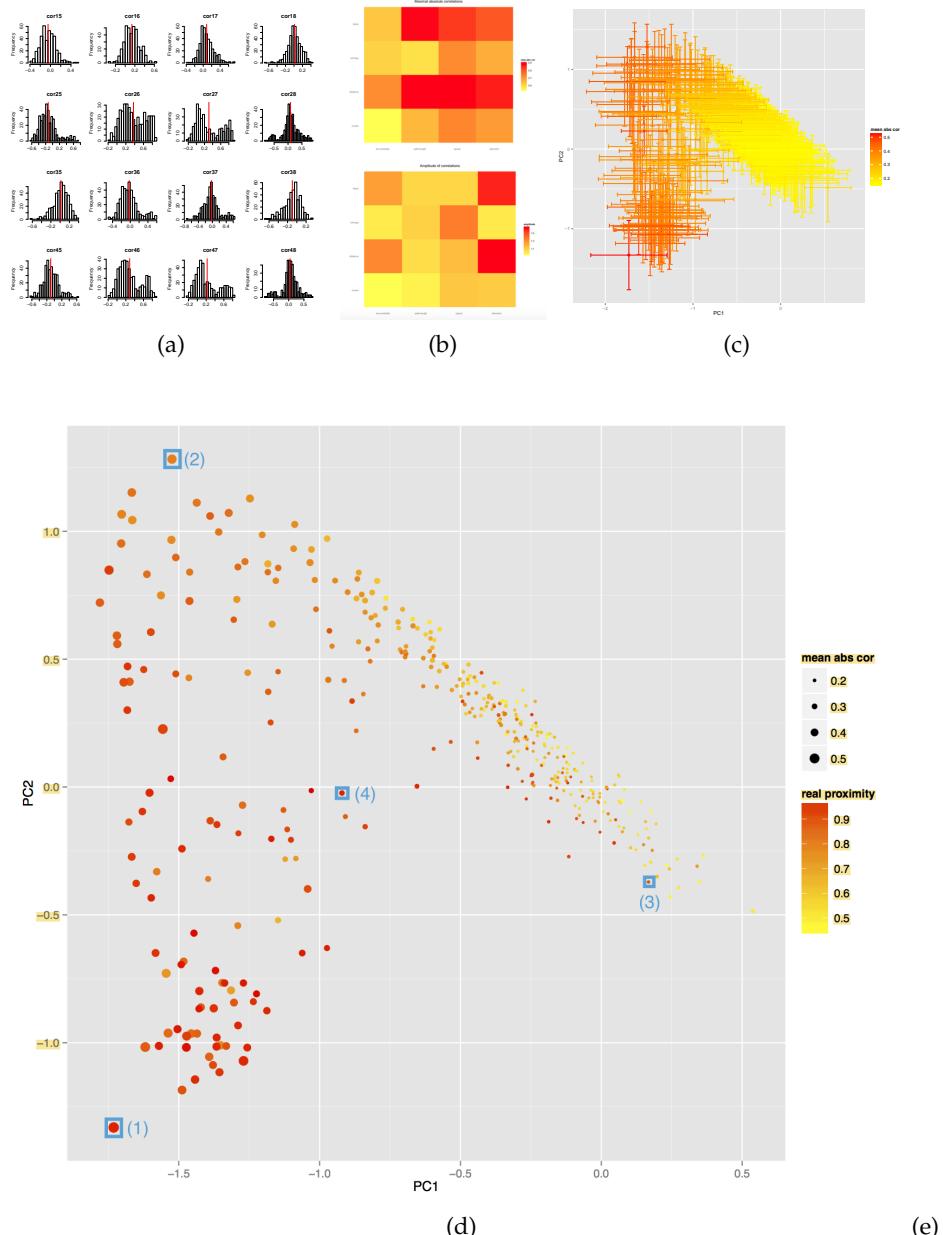


FIGURE 22 : C : (Florent) c'est bof comme ACP, il doit y avoir des modifications à faire en amont sur les données

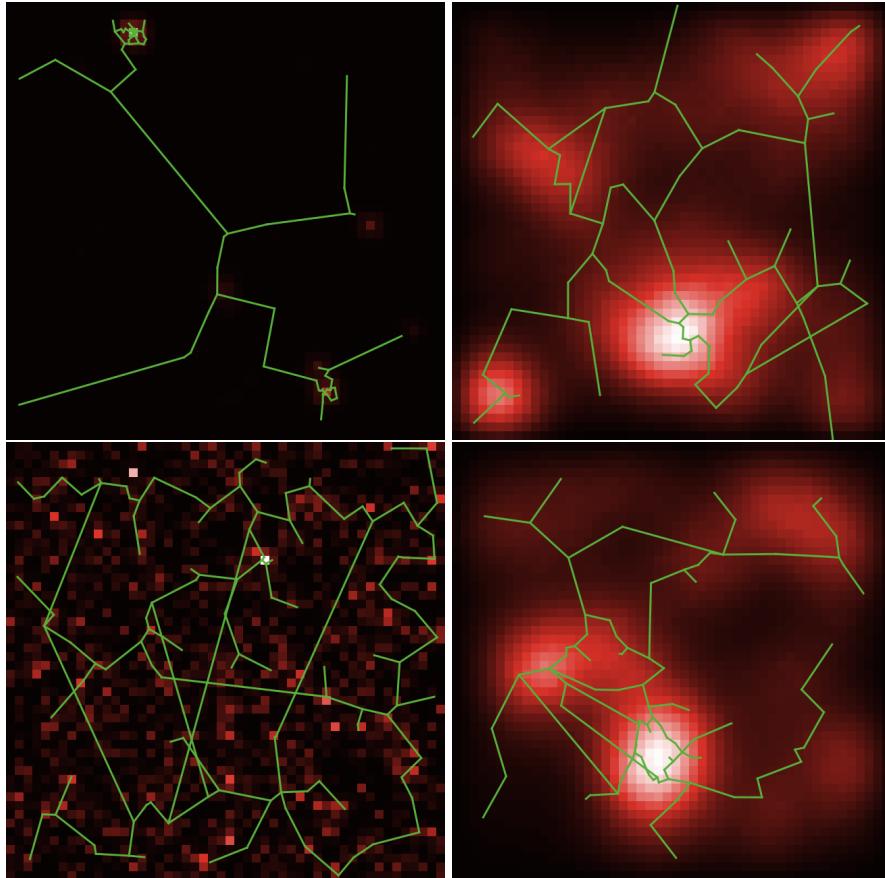


FIGURE 23 : Configurations obtenues pour les paramètres donnant les quatre points mis en évidence en 22 (d), dans l'ordre de gauche à droite et de haut en bas. Nous retrouvons des configurations de villes polycentriques (2 et 4), des établissements ruraux diffus (3) et une zone de densité agrégée faible (1). Se reporter à l'appendice ??

C : (Florent) il faut que tu discutes sur la forme de réseau obtenue et // réalité. qualitativement quels sont les aspects que tu t'attaches à reproduire ? Par ex. on trouve les différents cas suivants ; est ce que par exemple le modèle prévoit les bonnes "formes" selon différents ratios.

pour l'expérience. La figure ?? donne le détail des résultats de l'exploration. On retiendra les résultats marquants suivants au regard de la génération de données synthétiques corrélées :

- les distributions empiriques des coefficients de correlations entre indicateurs de forme et indicateurs de réseaux ne sont pas simples, pouvant être bimodales (par exemple $\rho_{46} = \rho[r, \bar{l}]$ entre l'index de Moran et le chemin moyen).
- On arrive à générer un assez haut niveau de correlation pour l'ensemble des indicateurs, la correlation absolue maximale variant entre 0.6 et 0.9; l'amplitude varie quant à elle entre 0.9 et 1.6, ce qui permet un large spectre de valeurs. L'espace couvert dans un plan principal a une étendue certaine mais n'est pas uniforme : on ne peut pas moduler à loisir n'importe quel coefficients, ceux-ci étant liés par les processus de génération sous-jacent. Une étude plus fine aux ordres suivants (correlation des correlations) serait nécessaire pour cerner exactement la latitude dans la génération.
- les points les plus corrélés en moyenne sont également ceux les plus proches des données réelles, ce qui confirme l'intuition d'une forte interdépendance en réalité.
- Des exemples concrets pris sur des points particuliers distants dans le plan principal montrent que des configurations de densité proches peuvent présenter des profils de correlations très différents.

Développements

Il est possible de raffiner cette étude en étendant la méthode de contrôle des correlations. La connaissance très fine du comportement de N (distribution statistiques sur une grille fine de l'espace des paramètres) conditionnée à D devrait permettre de déterminer exhaustivement $N^{<-1>}|D$ et avoir plus de latitude dans la génération des correlations. On pourra également appliquer des algorithmes spécifiques d'exploration pour essayer atteindre des configurations exceptionnelles réalisant un niveau de corrélation voulu, ou au moins pour découvrir l'espace des correlations atteignables par la méthode de génération [CHÉREL, COTTINEAU et REUILLOU, 2015].

6.3.2 Discussion

Positionnement Scientifique

Notre démarche s'inscrit dans un cadre épistémologique particulier. En effet, d'une part la volonté de multi-disciplinarité et d'autre part

l'importance de la composante empirique couplée aux méthodes d'exploration computationnelles, en font une approche typique des sciences de la complexité, comme le rappelle la structure de la feuille de route pour les systèmes complexes [BOURGINE, CHAVALARIAS et AL., 2009] qui croise des grandes questions transversales aux disciplines à une intégration verticale de celles-ci, qui implique la construction de modèles multi-échelles hétérogènes présentant souvent les aspects précédent. Le croisement de connaissances empiriques issues de la fouille de données avec celles issues de la simulation est souvent central dans leur conception ou leur exploration, et les résultats présentés ici en sont un exemple typique pour le cas de l'exploration.

Applications Directes

En partant du deuxième exemple, qui s'est arrêté à la génération des données synthétiques, on peut proposer des pistes d'application directe qui donneront un aperçu de l'éventail des possibilités.

- La calibration de la composante de génération de réseau, à densité donnée, sur des données réelle de réseau de transport (typiquement routier vu les formes heuristiques obtenues, il devrait par exemple être aisément d'utiliser les données ouvertes d'OpenStreetMap **C : (Florent) pas tant que cela (la couverture des données est bonne, après il peut y avoir des "encodages" bof (missing intersections e.g.)**³ qui sont de qualité raisonnable pour l'Europe, du moins pour la France [GIRRES et TOUYA, 2010], avec toutefois des ajustements à faire sur le modèle pour supprimer les effets de bord du à sa structure, par exemple en le faisant générer sur une surface étendue pour ne garder qu'une zone centrale sur laquelle la calibration aurait lieu) permettrait en théorie d'isoler un jeu de paramètres représentant fidèlement des situations existantes à la fois pour la forme urbaine et la forme du réseau. Il serait alors possible de dériver une "corrélation théorique" pour celles-ci, étant donné qu'une corrélation empirique n'est en théorie pas calculable puisqu'une seule instance des processus stochastiques est observée. Vu la non-ergodicité des systèmes urbains [PUMAIN, 2012b], il y a de fortes chances pour que ces processus soient différents d'une zone géographique à l'autre (ou selon un autre point de vue qu'ils soient dans un autre état des meta-paramètres, dans un autre régime) et que leur interprétation en tant que réalisations d'un même processus stochastique n'ait aucun sens, **C : (Florent) mais peut être que tu peux dégager des familles de situations selon des sous-processus dominants** entraînant l'impossibilité du calcul des covariations. En attribuant un jeu de données synthétiques similaire à une situation donnée, on serait capable

³ <https://www.openstreetmap.org>

de calculer une sorte de *correlation intrinsèque* propre à la situation, qui émerge en fait en réalité des interdépendances temporelles des composantes. Connaitre celle-ci renseigne alors sur ces interdépendances, et donc sur les relations entre réseaux et territoires.

- Comme déjà évoqué, la plupart des modèles de simulation nécessitent un état initial, généré artificiellement à partir du moment où la paramétrisation n'est pas effectuée totalement à partir de données réelles. Une analyse de sensibilité avancée du modèle implique alors un contrôle sur les paramètres de génération du jeu de données synthétique, vu comme méta-paramètre du modèle [COTTINEAU et al., 2015b]. Dans le cas d'une analyse statistique des sorties du modèle, on est alors capable d'effectuer un contrôle statistique au second ordre.
- On a étudié des processus stochastiques dans le premier exemple, au sens de séries temporelles aléatoires, alors que le temps ne jouait pas de rôle dans le second. On peut suggérer un couplage fort entre les deux composantes du modèle (ou la construction d'un modèle intégré) et observer les indicateurs et correlations à différents pas de temps de la génération. Dans le cas d'une dynamique, de par les rétroactions, on a nécessairement des effets de propagation et donc l'existence d'interdépendances décalées dans l'espace et le temps [PIGOZZI, 1980], étendant le domaine d'étude vers une meilleure compréhension des corrélations dynamiques.

Généralisation

On s'est limité au contrôle des premiers et second moments des données générées, mais il est possible d'imaginer une généralisation théorique permettant le contrôle des moments à un ordre arbitraire. Toutefois, la difficulté de génération dans un cas concret complexe, comme le montre l'exemple géographique, questionne la possibilité de contrôle aux ordres supérieurs tout en gardant un modèle à la structure cohérente au nombre de paramètres relativement faibles. Par contre, l'étude de structures de dépendances non-linéaires comme celles utilisées dans [CHICHEPORTICHE et BOUCHAUD, 2013] est une piste de développement intéressante.

6.3.3 *Conclusion*

On a ainsi proposé une méthode abstraite de génération de données synthétiques corrélées à un niveau contrôlé. Son implantation partielle dans deux domaines très différents montre sa flexibilité et l'éventail des applications potentielles. De manière générale, il

est essentiel de généraliser de telles pratiques de validation systématique de modèles par étude statistique, en particulier pour les modèles agents pour lesquels la question de la validation reste encore relativement ouverte.

CONCLUSION DU CHAPITRE

Troisième partie

SYNTHESIS : CONSTRUCTION OF CO-EVOLUTION MODELS

The buildings bricks, methods and tools are mainly set up for the culminating part of our work, which consists in the construction of models of co-evolution at different scales.

Processus
Attachement préférentiel
Diffusion/Etalement
Accessibilité
Gouvernance des Transports
Flux direct
Flux indirect/Effet tunnel <i>c'est le même processus, vu sous un angle différent : l'effet tunnel est l'absence d'un effet de proximité</i>
Centralité de proximité (accessibilité : généralisation)
Centralité de Chemin (correspond aux flux indirect : différents niveaux de généralité / sous-processus)
Proximité au réseau
Distance au centre (similar to agrégation?)

TABLE 4 :

TODO : à ce stade, expliquer lien entre les différents modèles : utiliser appendice unified framework urban growth

TODO : Plutôt en Conclusion Partie III : Towards operational Models : what is possible ; what is desirable ; etc.

C : faire le même tableau pour les modèles existants : vue plus large de l'ensemble des processus. pour chacun de ces modèles et de nos modèles, lister tous les processus potentiels ; faire une typologie ensuite. Q : typologie différente d'une pure empirique ? a creuser, et peut être intéressant dans le cadre du knowledge framework, comme illustration coevol connaissances.

ECHELLES ET PROCESSUS Partant des hypothèses tirées des enseignements empiriques et théoriques, on postulera *a priori* que certaines échelles privilégient certains processus, par exemple que la forme urbaine aura une influence au niveaux micro et mesoscopiques, tandis que les motifs émergeant des flux agrégés entre villes au sein d'un système se manifesteront au niveau macroscopique. Toutefois la distinction entre échelles n'est pas toujours si claire et certains processus tels la centralité ou l'accessibilité sont de bons candidats pour jouer un rôle à plusieurs échelles⁴ : il s'agira par la modélisation d'également tester ce postulat, par comparaison des processus nécessaires et/ou suffisants dans les familles de modèles à différentes échelles que nous allons mettre en place, en gardant à l'esprit des possibles développements vers des modèles multi-scalaires dans lesquels ces processus intermédiaires joueraient alors un rôle crucial.

⁴ on entend ici par "jouer un rôle" avoir une autonomie propre à l'échelle correspondante, c'est à dire qu'ils émergent *faiblement* des niveaux inférieurs.

VERS DES MODÈLES OPÉRATIONNELS DE COEVOLUTION

As previously stated, one of our principal aims is the validation of the network necessity assumption, that is the differentiating point with a classic evolutive urban theory. To do so, toy-model exploration and empirical analysis will not be enough as hybrid models are generally necessary to draw effective and well validated conclusions. We briefly give an overview of planned work in the following, that will be the conclusion of this Memoire.

OBJECTIFS

We expect to product *models of coevolution*, C : (Florent) expliciter la différence avec ce que tu as fait jusque là with the emphasis on processes of coevolution, to directly confront the theory. They will be necessary a flexible family because of the variety of scales and concrete cases we can include and we already began to explore in preliminary studies. Processes already studied can serve either as a thematic bases for a reuse as building bricks in a multi-modeling context, or as methodological tools such as synthetic data generator for synthetic control. Finally, we mean by operational models hybrid models, in the sense of semi-parametrized or semi-calibrated on real datasets or on precise stylized facts extracted from these same datasets. This point is a requirement to obtain a thematic feedback on geographical processes and on theory.

CAS D'ÉTUDE

Currently we expect to work on the following case studies to build these hybrid models :

- Dynamical data for Bassin Parisien should allow to parametrize and calibrate a model at this temporal and spatial scale.
- On larger scales, South African dataset of BAFFI will along empirical analysis also be used to parametrize hybrid co-evolution models.
- A possibility that is not currently set up (and that may however be difficult because of a disturbing closed-data policy among a frightening large number of scientists!) is the exploitation of French railway growth dataset (with population dataset) used in [BRETAGNOLLE, 2009a], that would also provide an interesting case study on other regimes, scales and transportation mode.

FEUILLE DE ROUTE

We give the following (non-exhaustive and provisory) roadmap for modeling explorations (theoretical and empirical domains being still explored conjointly) :

1. Complete the exploration of independent and weak coupled urban growth and network growth processes (all models presented in chapter ??), in order to know precisely involved mechanisms when they are virtually isolated, and to obtain morphogenesis scales.
2. Go further into the exploration of toy-model of non conventional processes such as governance network growth heuristic to pave the road for a possible integration of such modules in hybrid models.
3. Build a Marius-like generic infrastructure that implement the theory in a family of models that can be declined into diverse case studies.
4. Launch it and adapt it on these case studies.

Next steps would be too hypothetical if formulated, we propose thus to proceed iteratively in our construction of knowledge and naturally update this roadmap constantly.

- *La route est longue mais la voie est libre.*

7

CO-ÉVOLUTION À L'ECHELLE MACROSCOPIQUE

7.1 EXPLORATION DE SIMPOPNET

TODO : éventuellement ici pas seulement explorer simpopnet mais aussi le Portugali par exemple ?

7.1.1 Contexte

Quelle différentiel de connaissances obtenues peut s'observer, de la description conceptuelle ou thématique d'un modèle, à sa formalisation mathématique, son implémentation, son exploration systématique, jusqu'à son exploration approfondie à l'aide de meta-heuristiques spécifiques ? Notre postulat, qui découle à la fois de notre positionnement (voir [chapitre 3](#) sur la simulation) et d'expériences dont les modèles déroulés précédemment font partie, est que celui-ci est important, mais surtout de nature *qualitative*, c'est à dire que la nature même des connaissances subit des transitions abruptes lors de l'avancée de la démarche dans ce continuum. Le modèle SimpopNet introduit par [SCHMITT, 2014], qui est à notre connaissance l'unique modèle de co-évolution dans une perspective de la théorie évolutive des villes, est un exemple d'une telle démarche préliminaire qui nécessite d'être creusée, par exemple par l'exploration systématique.

7.1.2 Description du Modèle

7.1.3 Résultats

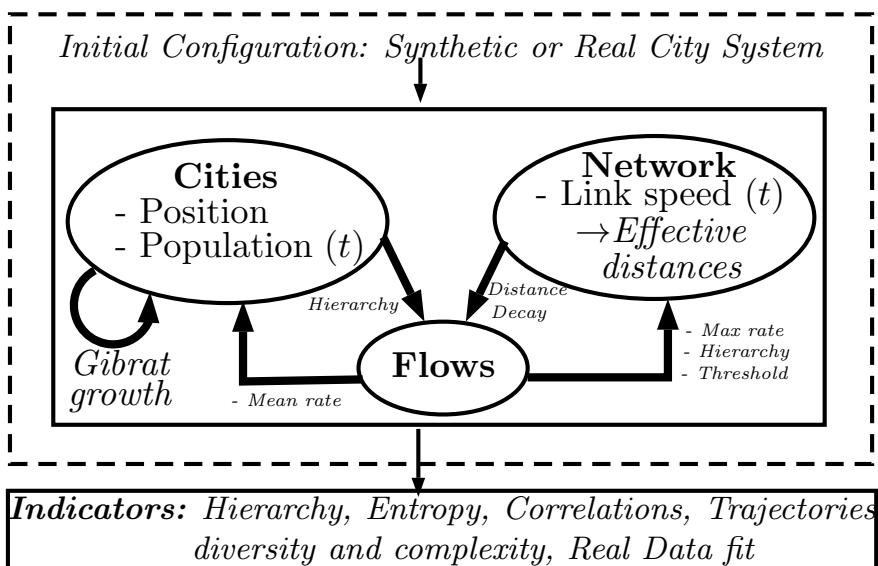


FIGURE 24 : Représentation abstraite des processus du modèle.

7.2 EXTENSION DYNAMIQUE DU MODÈLE D'INTERACTION

7.2.1 Modèle macroscopique de co-évolution

Hypothèses et choix de modélisation

Cette première approche se place dans une logique d'extension directe du modèle d'interactions au sein d'un système de villes présenté en chapitre ??, c'est à dire à une échelle macroscopique et avec une ontologie typique au systèmes de villes. Toujours dans un choix de simplicité, dont la relaxation sera explorée pour le cas d'application à la Chine avec l'ajout de variables économiques, nous restons ici à une description uni-dimensionnelle des villes par leur population. Concernant la croissance du réseau, nous proposons de se placer également à un niveau relativement agrégé et simplifié, en testant des heuristiques de croissance répondant à une demande, à différents niveaux d'abstraction.

Formulation Générique

7.2.2 Application à des Données Synthétiques

7.2.3 Applications à des cas d'étude

Système de Villes Français

DONNÉES DE RÉSEAU TODO : *Questions concrètes à poser au modèle, expériences ciblées : (i) le modèle calibre-t-il mieux les populations (en prenant en compte les paramètres supplémentaires);*

(ii) le modèle produit-il des formes de réseau crédibles?; (iii) éventuellement si les corrélations temporelles sont calculées sur les vraies données, le modèle peut-il être calibré au second ordre (sur les corrélations/causalités)?

C (JR) : attention, expliquer le choix des indicateurs de réseau, il faut qu'ils soient adaptés à l'échelle : cf Mimeur nombre d'intersection - relève un peu de la modélisation procédurale.

7.3 LE MODÈLE SIMPOPSINO

7.3.1 *Un modèle précurseur : SimpopJapan*

7.3.2 *Application du Modèle au Système de Villes Chinois*

TODO : application with HSR

7.3.3 *Vers le modèle SimpopSino*

C (JR) : justify why inclusion of economic variables is necessary for simpopsino; do some tests - if time, if not only model specification.

CONCLUSION DU CHAPITRE

★ ★

★

8

CO-EVOLUTION AT THE MESO-SCALE

Co-évolution à l'Echelle Mesoscopique

FIGURE 25 : C : (Florent) source ?

8.1 MODÈLES DE CROISSANCE DE RÉSEAU

8.1.1 Comparer les heuristiques de croissance de réseau

C : (Florent) cela doit remonter avant ton modèle : on est encore dans l'état de l'art Pour la croissance du réseau en tant que tel, de nombreuses heuristiques existent pour générer un réseaux sous certaines contraintes. Comme déjà développé précédemment, des modèles économiques de croissance de réseau au heuristiques d'optimisation locale, aux mécanismes géographiques ou à la croissance de réseau biologique, chacun a ses avantages et particularités propres. Un travail futur aura pour but de comparer ces diverses méthodes contre les valeurs réelles des indicateurs pour le réseau de routes européen. La Fig. 25 présente un travail préliminaire présenté dans [RAIMBAULT et GONZALEZ, May 2015] qui explore des applications des modèles de croissance de réseau biologique. D'autre part, comme présenté dans la section sur la reproductibilité, des modèles d'optimisation locale ont également été testés.

TODO : pour pouvoir comparer "toutes choses égales par ailleurs" les différentes heuristiques de génération de réseau, il est nécessaire des les explorer à densité fixée → cf thèse Mimeur et Morphogenesis ?

limitation du slime-mould [ADAMATZKY et JONES, 2010]

8.2 CO-EVOLUTION DES FORMES : INTERACTIONS ET MORPHOGENÈSE À L'ÉCHELLE MESOSCOPIQUE

8.2.1 *Modèle*

8.2.2 *Régimes de causalité*

8.2.3 *Calibration statique et dynamique*

8.3 MODÉLISATION DE LA GOUVERNANCE DU SYSTÈME DE TRANSPORT

Cette section fait un pas supplémentaire vers des modèles plus complexes. Un modèle jouet incluant des processus de gouvernance est décrit. Cette exploration répond de manière logique à notre cadre théorique et aux études précédentes, en particulier pour essayer de valider l'hypothèse de nécessité des réseaux : si des processus non-linéaires sont montrés nécessaires pour la validation sur des faits stylisés, cela pousse à argumenter pour sa validité.

8.3.1 *Contexte*

8.3.2 *Le Modèle Lutecia*

C : (Florent) on DC module : c'est à dire ? comparer quoi avec qui ?

C : (Florent) Implémentation : développer les aspects méthodo "techniques" ce n'est pas sale, au contraire

8.3.3 *Application au Delta de la Rivière des Perles*

CONCLUSION DU CHAPITRE

* *

*

CADRE THÉORIQUE

La théorie est un élément essentiel de toute construction scientifique, en particulier en Sciences Humaines pour lesquelles la définition des objets et questions de recherche sont plus ouverts mais aussi plus déterminants des directions de recherche alors prises. L'esprit de notre travail n'est pas de produire une théorie unifiée, mais des pistes pour des *Théories Intégrées*, c'est à dire s'appuyant sur une intégration horizontale et verticale au sens de la feuille de route [BOURGINE, CHAVALARAS et AL., 2009], mais aussi permettant une intégration des domaines de connaissance et une réflexivité, au sens qui seront précisés en section 9.3. Nous développons dans ce chapitre un cadre théorique à plusieurs niveaux. Il émerge naturellement de l'interaction des différentes composantes de la connaissance développées jusqu'ici. Dans sa partie thématique, il s'agit donc d'une clarification et unification d'hypothèse ainsi que de conclusions éparses.

Nous proposons d'abord de construire une *Théorie Géographique*, en quelque sorte un cadre théorique même si nous postulons qu'une Théorie propre a une plus grande portée de par son intégration forte avec les autres domaines de connaissance, qui fixera les objets étudiés et leur nature réelle (leur ontologie), ainsi que leur interrelations. Celle-ci permettra de produire des hypothèses précises qu'on cherchera à confirmer ou infirmer par la suite. Rester à un niveau thématique apparaît cependant ne pas être suffisant pour obtenir des lignes directrices générales sur le type de méthodologies et d'approches à utiliser. Plus précisément, même si certaines théories impliquent un usage plus naturel de certains outils¹, au niveau plus subtil de la mise en contexte au sens de l'approche prise pour implémenter la théorie (comme modèles ou analyses empiriques), la liberté de choix d'objets et d'approches en sciences sociales peut conduire à l'utilisation de techniques inappropriées ou des questionnements inadaptés (voir la section 3.2 pour l'exemple de l'usage inconsidéré des données massives et du calcul). Nous développons pour cela dans une seconde section (9.2) un cadre théorique à un niveau plus abstrait, visant à formaliser les entreprises de modélisation dans une certaine structure algébrique afin de capturer des articulations fondamentales entre diverses approches. Enfin, nous élaborons dans une dernière

¹ pour donner un exemple basique, une théorie mettant l'emphase sur la complexité des relations entre agents dans un système conduira généralement à utiliser de la modélisation basée agent et des outils de simulation, tandis qu'une théorie basée sur un équilibre macroscopique favorisera l'usage de dérivations mathématiques exactes.

section (9.3) un cadre de connaissances appliqué visant à expliciter des processus de production de connaissance sur les systèmes complexes. Celui-ci est illustré par une analyse fine de la genèse de la Théorie Evolutive des Villes, puis est ensuite appliqué de manière réflexive à l'ensemble de notre travail.

Ce chapitre sera éventuellement le plus délicat à la lecture, d'une part car il est fortement dépendant de la majorité des points thématiques traités précédemment et devrait être lu progressivement selon les concepts introduits (on touche encore aux limitations de la présentation linéaire), et d'autre part car les constructions théoriques introduites sont à un niveau d'abstraction progressif : en quelque sorte, chaque théorie est un cadre méta pour la précédente. On touche alors la question de la réflexivité, et dans quelle mesure celles-ci peuvent s'appliquer à elles-mêmes, en gardant à l'esprit que la séparation entre les niveaux n'est pas directement évidente : par exemple le cadre formel pour les systèmes socio-techniques pourrait être appliqué comme une formalisation du cadre de connaissances. Dans tous les cas, il faut comprendre la démarche à la fois comme une synthèse et comme une ouverture.

* * *

*

La première section de ce chapitre reprend un court passage de [RAIMBAULT, 2017a] ; la deuxième est entièrement inédite. La troisième a été proposée par [RAIMBAULT, 2017g] puis développée et appliquée dans [RAIMBAULT, 2017a], et son application réflexive a été présentée par [RAIMBAULT, 2017b].

9.1 UNE THÉORIE GÉOGRAPHIQUE DES TERRITOIRES ET DES RÉSEAUX

9.1.1 *Fondations*

Territoires Humains en Réseau

Notre premier pilier a déjà été construit précédemment lors de l'exploration thématique en Chapitre 1. Nous nous basons sur la notion de *Territoire Humain* élaborée par RAFFESTIN comme la base de la définition d'un système territorial. Elle permet de capturer les systèmes complexes géographiques humains dans l'ensemble de leur caractéristiques concrètes et abstraites, ainsi que dans leur représentations. Par exemple, un territoire métropolitain peut être appréhendé simplement par l'étendue fonctionnelle des flux pendulaires journaliers, ou par l'espace perçu ou vécu des différentes populations, le choix dépendant de la question précise à laquelle on cherche à répondre. Le territoire de RAFFESTIN devrait correspondre à un système cohérent de *synergetic inter-representation networks*, qui est à la fois une théorie et un modèle pour la cognition spatiale des individus et des sociétés, construite par *Portugali* et *Haken* (voir [PORTUGALI, 2011] pour une présentation synthétique). Elle postule que les représentations sont le produit du couplage fort entre les individus des cognitions et de leurs comportements individuels et collectifs. Cette approche au territoire est bien sûr un choix délibéré et que d'autres entrées, possiblement compatibles, peuvent bien sûr être prises [MURPHY, 2012]. Le ciment de ce pilier est renforcé par la théorie territoriale des réseaux de DUPUY, fournissant la notion de territoire humain en réseau, comme un territoire humain dans lequel un ensemble de réseaux transactionnels potentiels ont été réalisés, ce qui s'accorde par ailleurs avec les visions du territoire comme un lieu des réseaux [CHAMPOLLION, 2006]. Nous n'utiliserons pas les implications du développement de la notion de *lieu*, celles-ci étant trop éparses (voir définition de [Hypergeo]), et à cause de la redondance avec le territoire dans la vision de lien complexe entre représentation et réalité physique. Nous ferons pour ce premier pilier l'hypothèse fondamentale, déjà introduite en chapitre 1, que les réseaux réels sont des éléments nécessaires des systèmes territoriaux.

Théorie Evolutive des Villes

Le second pilier de notre construction théorique est la théorie évolutive des villes de PUMAIN, en relation étroite avec l'approche complexe que nous prenons de manière générale. Cette théorie a été introduite initialement dans [PUMAIN, 1997] qui argumente pour une vision dynamique des systèmes de ville, au sein desquels l'auto-organisation est essentielle. Les villes sont des entités spatiales évolutives inter-

dépendantes dont les interrelations font émerger le comportement macroscopique à l'échelle du système de villes. Le système de villes est aussi vu comme un réseau de villes, ce qui renforce sa vision en tant que système complexe. Chaque ville est elle-même un système complexe dans l'esprit de [BERRY, 1964], l'aspect multi-scalaire, au sens d'échelles autonomes mais ayant chacune un rôle spécifique dans les dynamiques du système, étant essentiel dans cette théorie, puisque les agents microscopiques véhiculent les processus d'évolution du système à travers des rétroactions complexes entre les échelles. Le positionnement de cette théorie au regard des Sciences des Systèmes Complexes a plus tard été confirmé [PUMAIN, 2003]. Il a été montré que la théorie évolutive fournit une interprétation des lois d'échelle qui sont omniprésentes dans les systèmes urbains, qui décleraient de la diffusion des cycles d'innovation entre les villes [PUMAIN et al., 2006], qui ont par ailleurs été mis en évidence de manière empirique pour plusieurs systèmes urbains [PUMAIN, PAULUS et VACCHIANI-MARCUZZO, 2009]. La notion de résilience d'un système de villes, induit par le caractère adaptatif des ces systèmes complexes, implique que les villes sont les moteurs et les adaptateurs du changement social [PUMAIN, 2010]. Enfin, la dépendance au chemin est source de non-ergodicité (voir définition en 4.1) au sein de ces systèmes, rendant les interprétations "universelles" des lois d'échelle développées par les physiciens incompatibles avec la théorie évolutive [PUMAIN, 2010]. La Théorie Evolutive des Villes a été élaborée conjointement avec des modèles de systèmes urbains : par exemple le modèle Simpop2 introduit par [BRETAGNOLLE, DAUDÉ et PUMAIN, 2006] est un modèle basé agent qui prend en compte des processus économiques, et simule sur de longues échelles de temps les motifs de croissance urbaine pour l'Europe et les Etats-unis [BRETAGNOLLE et PUMAIN, 2010b]. Les accomplissements les plus récents de la théorie évolutive reposent sur les productions du projet ERC GeoDiversity, présentées dans [PUMAIN et REUILLO, 2017], qui incluent de progrès avancés à la fois techniques (logiciel OpenMole² [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013]), thématiques (connaissance issue des modèles SimpopLocal [SCHMITT, 2014] et Marius [COTTINEAU, 2014]) et méthodologiques (modélisation incrémentale [COTTINEAU, CHAPRON et REUILLO, 2015]). Pour une analyse épistémologique par méthode mixte de la théorie évolutive, qui permet de renforcer cet aperçu bibliographique par une de sa genèse, en quelque sorte de sa *forme*, se référer à 9.3 qui l'utilise comme cas d'étude pour construire un cadre de connaissances. Ici, cette théorie nous permet d'interpréter les systèmes territoriaux comme systèmes complexes adaptatifs avec les implications listées ci-dessous.

² <http://openmole.org/>

Morphogenèse Urbaine

La notion de morphogenèse a été déjà explorée en profondeur et selon un point de vue interdisciplinaire en 6.1. Nous rappelons ici certains grands axes et dans quelles mesure ceux-ci contribuent à la construction de notre théorie. La morphogenèse a été particulièrement soulignée par TURING dans [TURING, 1952] lorsqu'il proposait d'isoler des règles chimiques élémentaires qui pourraient mener à l'émergence de l'embryon et à sa forme. La morphogenèse d'un système consiste en des règles d'évolution auto-cohérentes qui produisent l'émergence de ses états successifs, i.e. la définition précise de l'auto-organisation, avec la propriété supplémentaire qu'une architecture émergente existe, au sens de relations causales circulaires entre la forme et la fonction. Les progrès vers la compréhension de la morphogenèse de l'embryon (en particulier l'isolation de processus particuliers induisant la différentiation de cellules à partir d'une unique) sont relativement récents grâce à l'application des approches complexes en biologie intégrative [DELILE, DOURSAT et PEYRÉRAS, 2016]. Dans le cas des systèmes urbains, l'idée de morphogenèse urbaine, i.e. de mécanismes auto-cohérents qui produiraient la forme urbaine, est plutôt utilisé dans les champs de l'architecture et de l'urbanisme [HACHI, 2013] (comme e.g. la grammaire générative du "Pattern Language" d'ALEXANDER), en relation avec des théories de la forme urbaine [MOUDON, 1997]. Cette idée peut être poussée jusqu'à de très petites échelles comme celle du bâtiment [WHITEHAND, MORTON et CARR, 1999] mais nous l'utiliserons plus à une échelle mesoscopique, en termes de changements d'usage du sol à une échelle intermédiaire des systèmes territoriaux, avec des ontologies similaires à la littérature de modélisation de la morphogenèse urbaine (par exemple [BONIN et HUBERT, 2012] décrit un modèle de morphogenèse urbaine avec différentiation qualitative, tandis que [MAKSE et al., 1998] donne un modèle de croissance urbaine basé sur une distribution monocentrique de la population perturbée par des bruits corrélés). La notion de morphogenèse sera importante dans notre théorie en lien avec la modularité et l'échelle. La modularité d'un système complexe consiste en sa décomposition en sous-modules relativement indépendants, et la décomposition modulaire d'un système peut être vue comme un moyen de supprimer les corrélations non intrinsèques [KOLCHINSKY, GATES et ROCHA, 2015] (pour donner une image, penser à une diagonalisation par blocs d'un système dynamique du premier ordre). Dans le cadre de la conception et du contrôle de systèmes cyber-sociaux à grande échelle, des problèmes similaires surgissent naturellement et des techniques spécifiques sont nécessaires pour le passage à l'échelle des techniques simple de contrôle [WANG, MATNI et DOYLE, 2017]. L'isolation d'un sous-système fournit une échelle caractéristique correspondante. Isoler des processus de morphogenèse possibles implique une extraction

contrôlée (conditions au bord contrôlées par exemple) du système considéré, ce qui correspond à un niveau de modularité et donc à une échelle. Quand des processus auto-cohérents ne sont pas suffisants pour expliquer l'évolution d'un système (dans des variations raisonnables des conditions initiales), un changement d'échelle est nécessaire, causé par une transition de phase implicite dans la modularité. L'exemple de la croissance métropolitaine en est une très bonne illustration : la complexité des interactions au sein de la région métropolitaine sera croissante avec sa taille et la diversité des fonctions urbaines, ce qui conduit à un changement de l'échelle nécessaire pour comprendre les processus. L'émergence d'un aéroport international pourra dans certains cas influencer fortement le développement local, ce qui correspondra à une intégration significative dans un système plus vaste. Les échelles caractéristiques et la nature des processus pour lesquels ces changements ont lieu peuvent être des questions précisément approchées par l'angle de la modélisation. Il est important de noter qu'un sous-système territorial pour lequel la morphogenèse prend sens et dont les frontières sont bien définies peut être vu comme un *système auto-poiétique* au sens étendu de BOURGINE dans [BOURGINE et STEWART, 2004], i.e. comme un réseau de processus qui s'auto-reproduisent³ en régulant leur conditions aux bords, ce qui souligne la notion de frontière sur laquelle nous allons finalement nous attarder.

Co-évolution

Notre dernier pilier consiste en une clarification de la notion de *co-evolution*, sur laquelle HOLLAND apporte un éclairage pertinent à travers son approche des systèmes complexes adaptatifs (CAS) par une théorie des CAS comme agents dont la propriété fondamentale est de traiter des signaux grâce à leur frontières [HOLLAND, 2012]. Dans cette théorie, les systèmes complexes adaptatifs forment des agrégats à différents niveaux hiérarchiques, qui correspondent à différents niveaux d'auto-organisation, et les frontières sont intriquées horizontalement et verticalement de manière complexe. Cette approche introduit la notion de *niche* comme un sous-système relativement indépendant au sein duquel les ressources circulent (de la même façon que des communautés dans un réseau) : de nombreuses illustrations telles les niches écologiques ou économiques peuvent être données. Les agents au sein d'une niche sont dits en *co-évolution*. Empiriquement, les résultats obtenus témoignant d'une co-évolution à l'échelle mesoscopique comme en 4.2, confirment l'existence de niches pour certains aspects des systèmes territoriaux. La co-évolution implique ainsi de fortes interdépendances (impliquant des processus causaux circulaires) et une certaine indépendance au regard de l'extérieur de la niche. La

³ qui ne sont toutefois pas cognitifs, ne rendant pas ces systèmes morphogénétiques vivants au sens de auto-poiétique et cognitif

notion est naturellement flexible puisqu'elle dépendra des ontologies, de la résolution, des seuils, etc. que l'on considère pour définir le système. Nous postulons vu les indices d'existence obtenus dans les résultats empiriques, mais aussi les modèles reproduisant les processus de manière crédible sous une hypothèse d'isolation raisonnable, que ce concept peut se transmettre à la théorie évolutive urbaine et correspond à la notion de co-évolution décrite par PUMAIN : des agents co-évolutifs dans un système de villes consistent en une niche et ses flots, signaux et limites et sont donc des entités co-évolutives au sens de HOLLAND. Cette notion sera importante pour nous dans la définition des sous-systèmes territoriaux et de leur couplage. Nous gardons à l'esprit les potentialités et limitation du parallèle entre systèmes biologiques et systèmes sociaux décrits en 3.3.

9.1.2 *Synthèse : une théorie des systèmes territoriaux co-évolutifs en réseau*

Nous synthétisons les différents piliers en une théorie géographique autonome des systèmes territoriaux pour lesquels les réseaux jouent un rôle central pour la co-évolution des composantes du système. Pour les définitions des termes et les références, se référer à la section précédente. La formulation ici est voulue minimaliste.

Definition 1 - Système Territorial. *Un système territorial est un ensemble de territoires humains en réseau, c'est à dire des territoires humains au sein desquels et entre lesquels des réseaux réels existent.*

Le territoire est bien un élément du système territorial, qui de manière plus générale connecte différents territoires par les réseaux. A cette étape la complexité et le caractère évolutif et dynamique des systèmes territoriaux sont impliqués par les partis pris mais pas une partie explicite de la théorie. We supposerons pour simplifier une définition discrète des dimensions temporelles, spatiales et ontologiques, sous des hypothèses de modularité et de stationnarité locale. Cet aspect, à la fois pour le discret et la stationnarité, correspond à une simplification ontologique de la supposition d'une "échelle minimale" à laquelle les sous-systèmes fournissent une décomposition modulaire simple du système global. Elle reflète nos conclusions empiriques obtenues en Chapitre 5 et les modèles développés par la suite. On suppose également ergodicité locale, pour obtenir grâce à la démonstration proposée en 4.1 la propriétés de non-ergodicité globale typique des systèmes urbains.

Proposition 1 - Echelle discrètes. *Supposant une décomposition modulaire discrète d'un système territorial, l'existence d'un ensemble discret (τ_i, x_i) d'échelles temporelles et fonctionnelles pour le système territorial est équivalent à la stationnarité temporelle locale d'une spécification par système dynamique stochastique du système.*

Preuve (Tentative). Nous partons de l'hypothèse que tout système territorial peut être représenté par un ensemble de variables aléatoires, ce qui revient à avoir des objets et états bien définis et utiliser le Théorème de Transfert sur les événements des états successifs. Si $X = (X_j)$ est la décomposition modulaire, on a nécessairement quasi-indépendance des composantes au sens que $\text{Cov}[dX_j, dX_{j'}] \simeq 0$ à tout moment. Les transitions de stationnarité globales induisent des transitions dans chaque module, qui sont conservées si elles correspondent effectivement à un transition dans le sous-système. On obtient ainsi les échelles temporelles comme temps caractéristiques des sous-dynamiques. Les échelles fonctionnelles sont les étendues correspondantes dans l'espace d'état. ■

Cette proposition postule une représentation des dynamiques du système dans le temps. On peut noter que même en l'absence de représentation modulaire, le système dans son ensemble vérifiera la propriété. Cette définition des échelles permet d'introduire explicitement des boucles de rétroaction, puisqu'on peut par exemple conditionner l'évolution d'une échelle à celle d'une autre qui la contient, et ainsi l'émergence et la complexité, rendant la théorie compatible avec la théorie évolutive urbaine.

Assumption 1 - Imbrication des échelles et des sous-systèmes. Des réseaux complexes de retroaction existent à la fois entre et à l'intérieur des échelles [BEDAU, 2002]. De plus, un emboîtement horizontal et vertical des limites ne sera généralement pas hiérarchique.

Au sein de ces imbrications de sous-systèmes nous pouvons isoler des composantes en co-évolution en utilisant la morphogenèse. La proposition suivante est une conséquence de l'équivalence entre l'indépendance d'une niche et sa morphogenèse. La morphogenèse fournit la décomposition modulaire (sous hypothèse de stationnarité locale) nécessaire pour l'existence de l'échelle, donnant des sous-systèmes minimaux indépendants de manière verticale (échelle) et horizontale (espace).

Proposition 2 - Co-évolution des composantes. Les processus morphogénétiques d'un système territorial sont une formulation équivalente de l'existence de sous-systèmes co-évolutifs.

Nous formulons finalement la dernière hypothèse clé qui met les réseaux réels au centre des dynamiques co-évolutives, introduisant leur nécessité pour expliquer les processus dynamiques des systèmes territoriaux.

Assumption 2 - Nécessité des réseaux. L'évolution des réseaux ne peut pas être expliquée simplement par la dynamique des autres composantes territoriales et réciproquement, i.e. les sous-systèmes territoriaux co-évolutifs

contiennent les réseaux réels. Ceux-ci peuvent ainsi être à l'origine de changements de régime (transitions entre régimes stationnaires) ou de bifurcations plus conséquentes dans les dynamiques de l'ensemble du système territorial.

9.1.3 Contextualisation

Sur de longues échelles temporelles, une co-évolution globale a été montrée pour le systèmes ferroviaire français par [BRETAGNOLLE, 2009a]. A de plus petites échelles celle-ci est moins évidente (débat sur les effets structurants) mais nous supposons la présence d'effets co-évolutifs à toutes les échelles. Des exemples régionaux peuvent illustrer ce fait : Lyon n'a pas les mêmes relations dynamiques avec Clermont qu'avec Saint-Etienne, et la connectivité de réseau a probablement un rôle à y jouer (parmi les effets des dynamiques intrinsèques des interactions, et de la distance par exemple). A une plus petite échelle encore, nous partons du principe que les effets sont encore moins observables, mais précisément à cause du fait que la co-évolution est plus forte et les bifurcations locales se produisent avec une plus grande amplitude et une plus grande fréquence que dans les systèmes macroscopiques où les attracteurs sont plus stables et les échelles de stationnarité plus grandes. Nous pour cela que nous avons tenté d'identifier des bifurcations ou des transitions de phase dans des modèles jouets, des modèles hybrides, et des analyses empiriques, à différentes échelles, sur différents cas d'études et avec différentes ontologies.

Une difficulté dans notre construction est l'hypothèse de stationnarité locale, qui est essentielle pour formuler des modèles à l'échelle correspondante. Même si cela paraît une hypothèse raisonnable à plusieurs échelles et a déjà été observé des données empiriques [SANDERS, 1992], nous devrons le vérifier dans nos études empiriques. En effet, cette question est au centre des efforts de recherche courants pour appliquer les techniques d'apprentissage profond aux systèmes géographiques : BOURGINE a récemment développé un cadre pour extraire des motifs des systèmes complexes adaptatifs. En utilisant un théorème de représentation [KNIGHT, 1975], tout processus stationnaire discret est un *Modèle de Markov Caché*. Etant donné la définition d'un état causal comme $\mathbb{P}[\text{future}|A] = \mathbb{P}[\text{future}|B]$, la partition des états du système par la relation d'équivalence correspondantes permet de produire un *Réseau Récurrent* qui est suffisant pour déterminer l'état suivant du système, puisqu'il s'agit d'une fonction *déterministe* des états précédents et des états cachés [SHALIZI et CRUTCHFIELD, 2001] : $(x_{t+1}, s_{t+1}) = F[(x_t, s_t)]$. L'estimation des états cachés et de la fonction récurrente capture ainsi entièrement par apprentissage profond le comportement dynamique du système, i.e. l'information complète sur ses dynamiques et les processus internes. Les questions sont ensuite si les hypothèses de stationnarité peuvent être réglés par aug-

mentation des états du système, et si des données hétérogènes et asynchrones peuvent être utilisées pour initialiser des séries temporelles assez longues pour une estimation correcte du réseau de neurones ou de tout autre type d'estimateur. Ces questions sont reliées à l'hypothèse de stationnarité pour la première et à la non-ergodicité pour la seconde.

* * *

*

9.2 UN CADRE THÉORIQUE POUR L'ETUDE DES SYSTÈMES SOCIAUX-TECHNIQUES

Après avoir introduit une cadre théorique sur le plan thématique, nous développons un cadre plus général au sein duquel le précédent peut entrer. Il vise à contextualiser les directions générales de recherche à un niveau épistémologique mais formalisé, essayant d'obtenir une certaine structure algébrique pour capturer certaines propriétés des processus de modélisation.

9.2.1 Contexte

Contexte Scientifique

Les malentendus structurels entre les Sciences Sociales et Humanités d'une part, et les dénommées Sciences Exactes d'autre part, comme celui maintes fois évoqué déjà entre physiciens et géographes, loin d'être une règle nécessaire, semble toutefois avoir un impact conséquent sur la structure de la connaissance scientifique : [HIDALGO, 2015] montre comment la sociologie et la physique ont développé des méthodes d'analyse de réseau très similaire avec une inter-fertilisation faible. Ceux-ci peuvent être dus aux divergences épistémologiques qui elles-mêmes découlent de différences fondamentales dans les objets étudiés : les humains ne sont bien sûr pas des particules. Plus particulièrement, comme nous développons ici différents cadres théoriques, il est important de s'intéresser au rôle de celle-ci. La théorie, et en fait la signification elle-même du terme, a une place complètement différente dans l'élaboration de la connaissance, en partie à cause de différentes *complexités perçues*⁴ des objets étudiés. Par exemple, de nombreuses constructions mathématiques et par extension certaines en physique théorique sont *simples* au sens où elles sont résolubles de manière analytique (ou au moins semi-analytique)⁵, tandis que les sujets des Sciences Sociales tels les humains ou la société (pour prendre un exemple préconçu) sont *complexes* au sens de systèmes complexes. Cela implique un besoin accru d'une construction théorique (qui se base généralement sur l'empirique) pour identifier et définir qui sont nécessairement plus arbitraires dans la définition de leur limites, relations et processus, de par la multitude des points de vue possibles : PUMAIN suggère en effet dans [PUMAIN, 2005] une nouvelle approche de la complexité qui serait profondément ancrée dans les sciences sociales et qui serait "mesurée par la diversité des disciplines nécessaires pour élaborer une notion". Ces différences de fond sont natu-

⁴ Nous utilisons le terme *perçu* car la plupart des systèmes étudiés en physique peuvent être décrits comme simple alors qu'ils sont intrinsèquement complexe et finalement mal compris [LAUGHLIN, 2006].

⁵ nous prenons ici le parti que soluble analytiquement implique la simplicité, puisque le système n'exhibe alors pas d'émergence faible (voir 3.3).

rellement bénéfiques pour la diversité scientifique, mais les choses peuvent se corser quand les terrains d'étude se chevauchent, typiquement dans le cas de problématiques liées aux systèmes complexes comme déjà détaillé, comme l'exemple géographique des systèmes urbains a récemment montré [DUPUY et BENGUIGUI, 2015]. La Science des Systèmes Complexes⁶ est présentée par certains comme "un nouveau type de science" [WOLFRAM, 2002], et serait au moins symptomatique d'un changement de paradigme des pratiques, des approches analytiques "exactes" vers des approches computationnelles et *evidence-based* [ARTHUR, 2015], mais il est certain que cela permet de faire émerger, conjointement avec de nouvelles méthodologies, des nouveaux champs scientifiques au sens d'intérêts convergents de disciplines variées sur des questions transversales ou d'approches intégrées d'un champ particulier [BOURGINE, CHAVALARIAS et AL., 2009]. Notre travail s'ancre particulièrement dans ce cadre et n'aurait pas de sens s'il était déconnecté de ces aspects notamment computationnels (voir 3.2).

Objectifs

Dans ce contexte scientifique, l'étude de ce que nous désignons par *Systèmes socio-techniques*, que nous définissons de manière assez large comme des systèmes complexes hybrides qui incluent des agents ou objets sociaux qui interagissent avec des artefacts techniques et/ou un environnement naturel⁷, se situent précisément entre sciences sociales et sciences dures. L'exemple des systèmes urbains est relativement représentatif, puisque même avant l'arrivée de nouvelles approches prétendant être "plus exactes" que les approches des sciences sociales (typiquement par des physiciens, voir e.g. le positionnement de [LOUF et BARTHELEMY, 2014b], mais aussi par des chercheurs venant des sciences sociales comme BATTY [BATTY, 2013b]), une multitude d'aspects de l'étude des systèmes urbains étaient déjà traités dans des sciences dures très diverse, parmi lesquelles on peut citer sans hiérarchie particulière, l'hydrologie urbaine, la climatologie urbaine ou les aspects techniques des systèmes de transport, tandis que le centre de leur attention se reposait sur des sciences sociales comme la géographie, l'urbanisme, la sociologie, l'économie. D'où une place nécessaire de la théorie dans leur étude, vu son rôle comme domaine de connaissance pour la connaissance des systèmes complexes (voir le cadre introduit en 9.3).

⁶ que nous appelons délibérément ainsi même si des débats existent sur le fait de considérer comme une science en elle-même ou comme une façon différente de faire de la science.

⁷ les systèmes géographiques au sens de [DOLLFUS et DASTÈS, 1975] sont l'archetype de tels systèmes, mais cette définition peut couvrir d'autres types de systèmes comme un système de transport étendu, des systèmes sociaux pris dans un contexte environnemental, des systèmes industriels compliqués considérés avec leur utilisateurs, etc.

Nous proposons dans cette section de construire une théorie, ou plutôt un cadre théorique, pour faciliter certains aspects de l'étude de tels systèmes. De nombreuses théories existent déjà dans l'ensemble des champs liés à ce type de questionnement, et aussi à de plus haut niveaux d'abstraction concernant des méthodes comme e.g. la modélisation basée agent, mais il n'existe à notre connaissance pas de cadre théorique qui incluraient l'ensemble des points suivants que nous jugeons cruciaux (et qui peuvent être compris comme une base informelle de notre théorie) :

1. une définition précise et une emphase particulière sur la notion de couplage entre sous-systèmes, en particulier permettant de qualifier ou quantifier un certain niveau de couplage : dépendance, interdépendance, etc. entre composantes.
2. une précise définition de l'échelle, incluant l'échelle temporelle et l'échelle pour d'autres dimensions.
3. en conséquence des points précédents, une définition précise de ce qu'est un système.
4. la prise en compte de la notion d'émergence pour capturer les aspects multi-scalaires des systèmes.
5. une place centrale de l'ontologie dans la définition des systèmes, i.e. du sens dans le monde réel donné à ses objets⁸.
6. la prise en compte d'aspects hétérogènes du même système, qui peuvent être des composantes hétérogènes mais aussi différents points de vue sur le système qui se complètent.

La suite de cette section est organisée de la façon suivante : nous construisons la théorie dans la sous-section suivante en restant à un niveau abstrait, et proposons une première application à la question des sous-systèmes co-évolutifs. Nous discutons ensuite le positionnement au regard de théories existantes, ainsi que les développements possibles et des applications concrètes.

9.2.2 Construction de la Théorie

Perspectives et Ontologies

Le point de départ pour construire la théorie est une approche épistémologique perspectiviste des systèmes introduite par GIERE [GIERE, 2010c]. Pour résumer, cette position interprète toute démarche scientifique comme une perspective, au sein de laquelle chacun poursuit

⁸ comme déjà expliqué précédemment, ce positionnement combiné à l'importance de la structure pourrait être relié au *Réalisme Structurel Ontologique* dans des approfondissements.

certains objectifs et utilise ce qui est appelé *un modèle* pour les atteindre. Le modèle n'est alors rien de plus qu'un medium scientifique. VARENNE a développé [VARENNE, 2010a] une typologie fonctionnelle des modèles qui peut être interprété comme un raffinement de cette théorie. Relâchons dans un premier temps cette précision potentielle et utilisons les perspectives comme des approximations des objets et concepts indéfinis. En effet, diverses visions du même objet (pouvant être complémentaires ou divergentes) ont la propriété de partager au moins l'objet lui-même, d'où notre proposition de définir les objets (et plus généralement les systèmes) à partir d'un ensemble de perspectives sur ceux-ci, qui vérifient certaines propriétés que nous formalisons par la suite.

Une perspective est définie dans notre cas comme une *Dataflow Machine M* au sens de [GOLDEN, AIGUIER et KROB, 2012], que nous considérons comme une boîte noire transformant un flux de données d'entrée en flux de sortie à une échelle de temps associée, et qui correspond au model comme medium. Celle-ci fournit un moyen adapté de représenter un modèle et d'y associer échelle de temps et données. On y associe un ontologie O au sens de [LIVET et al., 2010], i.e. un ensemble d'éléments qui correspondent à une entité (qui peut être un objet, un agent, un processus, un état, un concept, c'est à dire tout élément modulaire formalisable) du monde réel. Nous incluons seulement ces deux aspects (le modèle et les objets représentés) de la théorie de Giere, en faisant l'hypothèse que le but et le producteur de la perspective sont en fait contenus dans l'ontologie s'ils font sens pour l'étude du système : par exemple, dans le cas des sondages subjectifs en anthropologie ou sociologie, le sondeur est un élément clé est sera nécessairement inclus dans l'ontologie. De même pour l'objectif poursuivi, tout particulièrement en sciences humaines où la recherche n'est jamais neutre comme nous l'avons vu en 3. Formalisons cette définition :

Definition 2 *Une perspective sur un système est donnée par une Dataflow Machine M = (i, o, T) et une Ontologie associée O. Nous supposons que l'ontologie peut être décomposée de manière discrète en éléments atomiques O = (O_j)_j.*

Les éléments atomiques de l'ontologie peuvent être des constituants particuliers du systèmes, comme des agents ou des composantes, mais aussi des processus, interactions, états ou concepts par exemple. L'ontologie peut être vue comme la description exhaustive et rigoureuse du contenu de la perspective. L'hypothèse d'une *Dataflow Machine* implique que les entrées et sorties potentielles peuvent être quantifiées, ce qui n'est pas nécessairement restrictif aux perspectives quantitatives, puisque la plupart des approches qualitatives peuvent être traduites en variables discrètes à partir du moment où l'ensemble des possibles est connu ou supposé.

Nous définissons alors le système de manière "réciproque", i.e. à partir d'un ensemble de perspectives sur ce qui constitue alors le système :

Definition 3 *Un système est un ensemble de perspectives sur un système : $S = (M_i, O_i)_{i \in I}$, où I n'est pas nécessairement fini.*

Nous désignons par $\mathcal{O} = (O_{j,i})_{j,i \in I}$ l'ensemble des éléments dans les ontologies.

Comme on part des perspectives sur un système pour définir le système dans son ensemble, il n'y a pas de contradiction. On peut noter qu'à ce stade de la construction, il n'existe pas nécessairement de cohérence structurelle, au sens d'une correspondance avec une structure réelle, sur ce qu'on appelle un système, puisque étant donné notre définition très large nous pourrions par exemple considérer un système comme une perspective sur un véhicule conjointement à une perspective sur un système de villes, ce qui ne fait pas raisonnablement sens. Des définitions approfondies et développements doivent permettre de se rapprocher des définitions classiques d'un système (entités en interaction, artefacts précisément définis, etc.). De la même manière, la définition d'un sous-système sera donnée plus loin. Les éléments de l'approche déjà introduits permettent jusqu'ici de répondre aux points trois, cinq et six des recommandations.

PRÉCISION SUR L'ASPECT RÉCURSIF DE LA THÉORIE Une conséquence directe de ces définitions doit être détaillée : le fait qu'elles peuvent être appliquées de manière récursive. En effet, on peut imaginer prendre comme perspective un système dans notre sens, c'est à dire un ensemble de perspectives sur un système, et le faire à tout ordre. Si on considère un système à n'importe quel sens classique, alors le premier ordre peut être interprété comme une épistémologie du système, i.e. l'étude de perspectives sur un système. Une ensemble de perspectives sur des systèmes en relation peut sous certaines conditions être un domaine ou un champ d'étude, et donc un ensemble de perspectives sur diverses perspectives l'épistémologie d'un champ. On peut proposer des analogies supplémentaires pour traduire l'idée derrière le caractère récursif de la théorie. C'est en effet crucial pour la signification et la cohérence de la théorie, notamment pour les raisons suivantes : (i) le choix des perspectives qui constituent un système est nécessairement subjectif et peut donc être compris comme une perspective en lui-même, et ainsi une perspective sur un système si l'on est en mesure de construire une ontologie générale; (ii) nous utiliserons des relations entre ontologies par la suite, dont la construction est basée sur l'émergence est également subjective et vue comme perspectives. Ces aspects de réflexivité sont fondamentaux, en écho à la discussion de 3.3 sur la production de connaissance et la nature de la complexité.

Graphe Ontologique

Nous proposons ensuite la structure du système en reliant les ontologies. Cette approche pourrait éventuellement être mise en perspective par rapport à un positionnement épistémologique de réalisme structurel [FRIGG et VOTSI, 2011], c'est à dire que les théories tendent à capturer une certaine structure existante du monde réel, puisqu'une connaissance du monde est ici partiellement contenue dans la structure des modèles, tout en gardant à l'esprit que notre position s'en éloigne en partie de par la conjugaison des perspectives qui induit un certain "degré de constructivisme" comme expliqué en 3.3. Pour cette raison, nous faisons le choix d'appuyer le rôle de l'émergence, suivant l'intuition qu'il pourrait s'agir d'un outil pratique minimaliste pour capturer de façon raisonnable la structure d'un système complexe⁹. Nous prenons pour cet aspect le positionnement de BEAU sur les différents types d'émergence déjà présenté plusieurs fois, en particulier sa définition de l'émergence faible donnée dans [BEAU, 2002]. Rappelons brièvement les définitions que nous utiliserons par la suite. BEAU commence par définir les propriétés émergentes puis étend le concept aux phénomènes, entités, etc. De la même manière, notre cadre n'est pas restreint aux objets ou propriétés et inclut ainsi les définitions généralisées comme lien entre ontologies. Nous appliquons la notion d'émergence sous les deux formes suivantes¹⁰ :

- *Emergence nominale* : une ontologie O' est inclue dans une autre ontologie O mais l'aspect de O qui est dit nominalement émergent en rapport à O' ne dépend pas de O' .
- *Emergence faible* : une partie d'une ontologie O peut être dérivée de manière computationnelle par agrégation et interactions entre les éléments d'une ontologie O' .

Comme développé précédemment, la présence d'émergence, et spécifiquement d'émergence faible, constitue une perspective en soi. Elle peut être conceptuelle et postulée comme un axiome dans une théorie thématique, mais aussi expérimentale si des traces d'émergence faible sont effectivement mesurées entre objets. Dans tous les cas, la relation entre ontologies doit être encodée dans une ontologie, ce qui n'était pas nécessairement introduit dans la définition initiale d'un système. Ainsi pour simplifier, les perspectives permettent de décomposer le système en briques ontologiques spécifiant une description "complète".

⁹ ce qui bien sûr ne peut être formulé comme une affirmation prouvable car cela dépendra de la définition d'un système, etc.

¹⁰ la troisième forme rappelée par BEAU, *l'émergence forte*, ne sera pas utilisée, car nous avons besoin de capturer rien de plus des relations de dépendance et d'autonomie, et l'émergence faible est plus adéquate en termes de systèmes complexes, puisqu'elle n'assume pas "des pouvoirs causaux irréductibles" aux objets des échelles supérieures à un niveau donné. L'émergence nominale est utilisée pour capturer des relations d'inclusion entre les ontologies.

Nous faisons pour cette raison l'hypothèse suivante importante par la suite :

Assumption 3 *Un système peut être partiellement structuré par son extension avec une ontologie qui contient (pas nécessairement uniquement) des relations entre les éléments des ontologies de ses perspectives. Nous la désignons ontologie de couplage et supposons son existence par la suite. Nous postulons de plus son atomicité, i.e. si O est en relation avec O' , alors tout sous-ensemble de O, O' ne peuvent être en relation, ce qui n'est pas contraignant puisqu'une décomposition en des sous-ensembles indépendants assurera cette propriété si elle n'est pas vérifiée initialement.*

Cette hypothèse revient concrètement qu'il est possible de coupler des perspectives, c'est à dire souvent des modèles en pratique, et que ce couplage peut être représenté de façon similaire. Notre expérience pratique du couplage tout au long de nos travaux nous pousse à faire cette hypothèse : tant que les systèmes considérés sont "raisonnables" (choisi raisonnablement l'un par rapport à l'autre, et donc choisi pour être couplés en quelque sorte), il est toujours possible de les coupler.

Cela nous permet d'exhiber des relations d'émergence pas seulement au sein d'une perspective elle-même, mais également entre les éléments de différentes perspectives. Nous définissons ensuite des relations de pré-ordre entre les sous-ensemble des ontologies :

Proposition 3 *Les relations binaires suivantes sont des pré-ordres sur $\mathcal{P}(O)$:*

- *Emergence (basée sur l'émergence faible) : $O' \preceq O$ si et seulement si O émerge faiblement de O' .*
- *Inclusion (basée sur l'émergence nominale) : $O' \Subset O$ si et seulement si O émerge nominalement de O' .*

Avec la convention qu'il peut être admis qu'un objet émerge de lui-même, on a réflexivité (si une telle convention paraît absurde, on peut définir les relations comme O émerge de O' ou $O = O'$). La transitivité est clairement contenue dans la définition de l'émergence.

Notons que la relation d'inclusion est plus général qu'une inclusion entre ensembles, puisqu'elle traduit une inclusion "au sein" des éléments de l'ontologie. Par exemple, une ontologie peut supposer un couplage fort non-décomposable (qui serait une hypothèse de la perspective en elle-même), et une autre perspective contenir l'un des éléments de ce couplage. Nous allons voir que ces relations d'ordre vont nous permettre de définir un graphe par l'algorithme de réduction qui suit.

Definition 4 *Le graphe ontologique est construit par induction de la manière suivante :*

1. *Un graphe est construit, avec pour noeuds des éléments de $\mathcal{P}(\mathcal{O})$ et des liens de deux types : $E_W = \{(O, O') | O' \preccurlyeq O\}$ et $E_N = \{(O, O') | O' \sqsubseteq O\}$*
2. *Les noeuds sont réduits¹¹ par : si $o \in O, O'$ et $(O' \preccurlyeq O$ ou $O' \sqsubseteq O)$ mais pas $(O \preccurlyeq O'$ or $O \sqsubseteq O')$, alors $O' \leftarrow O' \setminus o$*
3. *Les noeuds avec des ensemble se recouplant sont fusionnés, en gardant les liens liant des noeuds fusionnés. Cette étape assure des noeuds ne se recouplant pas.*

Arbre Ontologique Minimal

La structure topologique du graphe, qui contient en un sens la *structure du système*, peut être réduite en un arbre minimal qui capture la structure hiérarchique essentielle pour la théorie.

Nous devons d'abord donner cohérence au système :

Definition 5 *Une partie cohérente du graphe ontologique est une composante du graphe faiblement connectée au sens d'un graphe dirigé. Nous assumons pour la suite travailler sur une partie cohérente.*

La notion de système cohérent, ainsi que de sous-système ou d'échelle de temps des noeuds qui seront définies par la suite, nécessite de reconstruire des perspectives à partir des éléments ontologiques, i.e. l'opération inverse de ce qui a été fait dans notre procédure qui peut être vue comme une deconstruction.

Assumption 4 *Il existe $\mathcal{O}' \subset \mathcal{P}(\mathcal{O})$ tel que pour tout $O \subset \mathcal{O}'$, il existe une Dataflow Machine M correspondante telle que la perspective correspondante est cohérente avec les éléments initiaux du système (i.e. les machine sont équivalentes sur les parties communes des ontologies). Si $\Phi : M \mapsto O$ est la correspondance initiale, nous notons cette construction réciproque étendue par $M' = \Phi^{<-1>}(O)$.*

REMARQUE Cette hypothèse pourrait éventuellement être changée en une proposition prouvable, en supposant que l'ontologie de couplage correspond effectivement à une perspective de couplage, dont la composante *Dataflow Machine* est cohérente avec les entités couplées. Ainsi, le postulat de décomposition de [GOLDEN, AIGUIER et KROB, 2012] devrait permettre d'identifier des composantes de base correspondantes à chaque élément de l'ontologie, et construire ainsi la nouvelle perspective par induction. Nous trouvons toutefois ces hypothèses trop restrictives, puisque par exemple divers éléments de l'arbre ontologique peuvent être modélisés par la même machine irréductible, à l'image d'une équation différentielle aux variables agrégées. Nous préférons être moins restrictifs et postuler l'existence de la

¹¹ la procédure de réduction vise à supprimer la redondance, gardant une entité au plus haut niveau où elle existe.

correspondance inverse sur certaines sous-ontologies, qui devraient être en pratique celles sur lesquelles le couplage peut effectivement être modélisé.

Grace à l'hypothèse ci-dessus, on peut définir le système cohérent comme l'image réciproque de la partie cohérente du graphe ontologique. Cela permet la connectivité du système qui est un pré-requis pour la construction de l'arbre.

Proposition 4 *La décomposition arborescente du graphe ontologique dans laquelle les noeuds contiennent les composantes fortement connexes est unique. L'arbre réduit, qui correspond au graphe ontologique les composantes fortement connexes ont été fusionnées et les liens gardés, est nommé Arbre Ontologique Minimal.*

Proof (esquisse) L'unicité découle de la définition univoque puisque les noeuds sont fixés comme les composantes fortement connexes. Il s'agit trivialement d'une décomposition en arbre puisque dans un graphe dirigé, les composantes fortement connexes ne se recoupent pas, d'où la cohérence de la décomposition.

Toute boucle $O \rightarrow O' \rightarrow \dots \rightarrow O$ dans le graphe ontologique suppose que tous ses éléments sont équivalents au sens de \preccurlyeq . Ces boucles d'équivalence devrait aider à définir la notion de couplage fort comme une application de la théorie, avec cependant un caractère qualitatif dans la nature du couplage, ne permettant pas une définition fine de la force de couplage par exemple.

L'Arbre Minimal Ontologique (MOT) est un arbre au sens non-dirigé, mais une forêt au sens dirigé. Sa topologie contient une représentation des hiérarchies du système. Les sous-systèmes cohérents sont définis à partir de l'ensemble \mathcal{B} des branches de la forêt, comme $(\Phi^{<-1>}(\mathcal{B}), \mathcal{B})$. L'échelle de temps d'un noeud, et par extension d'un sous-système, est l'union est échelles de temps des machines correspondantes. Les niveaux de l'arbre sont définis à partir des noeuds racine, et les relations d'émergence entre les noeuds implique une inclusion verticale entre échelles de temps.

Action sur des Données

De la même manière que les actions de groupes permettent de donner structure à l'utilisation d'un groupe sur un ensemble (généralement de données), une piste de développement puissante serait l'ajout à la théorie de l'aspect essentiel de relation à la réalité par une action des noeuds de l'arbre ontologique sur des ensembles de données. Cette opération est hors de propos pour l'instant car nous n'avons pas encore exploité la structure interne des *dataflow machines*. Une piste, que nous confirmions comme ouverture dans la section suivante 9.3, impliquerait le couplage de ce cadre avec le cadre de connaissances qui y est introduit.

Echelles

Enfin, nous proposons de définir les échelles associées à un système. Suivant [MANSON, 2008], un continuum épistémologique de visions sur l'échelle est une conséquence des différences propres à chaque discipline, comme nous avons développé en introduction. Cette proposition est en fait compatible avec notre cadre, puisque la construction d'échelles pour chaque niveau de l'arbre ontologique résulte en une grande variété d'échelles.

Soit (M, O) un sous-système et T l'échelle de temps correspondante. Nous proposons de définir "l'échelle thématique" (par exemple l'échelle spatiale) en supposant un théorème de représentation, i.e. qu'un aspect (aspect thématique) de la machine peut être représenté par une variable d'état dynamique $\vec{X}(t)$. Etant donné un opérateur d'échelle¹² $\|\cdot\|_s$ et que la variable d'état est différentiable à un certain niveau, *l'échelle thématique* pour cet aspect, c'est à dire l'échelle typique à laquelle les agents ou processus correspondants opèrent (pouvant être multiple si l'opérateur est multidimensionnel), est définie par $\|(d^k \vec{X}(t))_k\|_s$.

9.2.3 Applications et discussion

Le cas particulier des systèmes géographiques

Dans [DOLLFUS et DASTÈS, 1975], DURAND-DASTÈS introduit une définition des systèmes et structures géographiques, la structure étant le contenant spatial des systèmes vus comme des systèmes complexes ouverts en interaction (donné par ses éléments et leur attributs, les relations entre éléments et les entrée/sorties avec le monde extérieur). Pour un système donné, sa définition est une perspective, complétée par la structure pour avoir un système selon notre sens. Selon la manière dont les relations sont définies, cela peut être plus ou moins aisément d'extraire la structure ontologique.

Modularité et sous-systèmes en co-évolution

Pour l'exemple des systèmes urbains, la théorie évolutive des villes entre dans ce cadre en utilisant notre théorie thématique développée dans la section précédente. La décomposition en sous-systèmes décorrélés fournit précisément des composantes fortement couplées comme des composantes en co-évolution. La corrélation entre sous-systèmes devrait d'une certaine façon être corrélée à la distance topologique dans l'arbre. Si on définit les éléments d'un noeud avant réduction comme *éléments fortement couplés*, dans le cas d'ontologies

¹² qui peut être de nature variée : étendue, étendue probabiliste, échelles spectrales, échelles de stationnarité, etc.

dynamiques, cela fournit une définition de la *co-évolution* et de sous-systèmes en co-évolution, équivalente à la définition thématique.

Discussion

LIEN AVEC DES CADRES EXISTANTS Un lien avec le cadre de Cottineau-Chapron pour la multi-modélisation [CHÉREL, COTTINEAU et REUILLOU, 2015] pourrait être fait dans le cas où ils ajouteraient la couche bibliographique, qui correspondrait à la reconstruction des perspectives. [REYMOND et CAUVIN, 2013] propose la notion de “couplage interdisciplinaire” qui est proche de notre notion de coupler des perspectives. Une correspondance avec les approches de Système de Systèmes (voir e.g. [LUZEAUX, 2015] pour un cadre récent englobant la modélisation et la description des systèmes) pourrait être également possible puisque nos perspectives sont construites comme des *Dataflow Machines*, mais avec la différence cruciale que la notion d'émergence est centrale dans notre cas.

CONTRIBUTION À L'ÉTUDE DES SYSTÈMES COMPLEXES Nous ne prétendons pas exhiber une théorie des systèmes (il faut généralement se méfier de la cybernétique, la systémique etc. qui ne peuvent pas tout modéliser), mais plutôt un cadre majoritairement axiomatique et la structure associée pour guider les questions de recherche (e.g. dans notre cas les conséquences directes sont les études d'épistémologie quantitative qui vient de la construction des systèmes comme perspectives ; les études empiriques pour construire des ontologies robustes pour les perspectives ; des études thématiques ciblées pour révéler des relations causales ou l'émergence pour la construction des réseaux ontologiques ; l'étude des couplages comme processus contenant possiblement de la co-évolution ; l'étude des échelles ; etc.). Cela peut être compris comme une meta-théorie dont l'application donne une théorie, la théorie thématique qui précède étant une implémentation aux systèmes territoriaux en réseau. Nous appuyons la notion de système socio-technique, croisant une approche des systèmes sociaux complexes (ontologies) avec une description des artefacts techniques (*Dataflow Machines*), prenant “le meilleur des deux mondes”.

Réflexivité

Nous pouvons tirer de l'application de ce cadre à notre travail, c'est à dire d'une réflexivité, une clarification des directions de recherche menées jusqu'ici, et donc de la co-construction des réponses à ces questions avec les différents cadres théoriques.

1. L'approche perspectiviste implique une compréhension large des perspectives existantes sur un système, et des possibilités de couplage entre celle-ci ; d'où une emphase sur l'épistémologie quantitative qui inclue la revue systématique algorithmique

(exploration de l'espace des connaissances), la cartographie des connaissances (extraction de sa structure) et de possibilités de fouille de contenu (raffinement au niveau atomique de la connaissance scientifique) qui correspondent au travail de 2.2.

2. A un niveau plus fin de particularité, la connaissance des perspectives signifie une connaissance des faits stylisés empiriques, comme par exemple ceux pour le traffic routier 5.1, les prix des carburants 5.2, les formes urbaines et de réseau 4.1.

★ ★

★

9.3 UN CADRE DE CONNAISSANCES APPLIQUÉ POUR L'ETUDE DES SYSTÈMES COMPLEXES

La complexité de la production de connaissance sur des systèmes complexes est bien connue, mais il n'existe toujours pas de cadre de connaissance qui rendrait à la fois compte d'une certaine structure de la production de connaissance à un niveau épistémologique et serait directement applicable à l'étude et au management des systèmes complexes. Nous posons ici les bases d'un tel cadre, en commençant par analyser en détail l'étude de cas de la construction d'une théorie géographique des systèmes territoriaux complexes, au travers de méthodes mixtes, plus précisément des analyses qualitatives d'entretiens et une analyse quantitative de réseau de citation. Nous pouvons par cela construire de manière inductive un cadre qui considère les entreprises de production de connaissance comme des perspectives, dont les composantes sont en co-évolution au sein de domaines de connaissances complémentaires. Nous discutons finalement des applications et développements potentiels.

La compréhension des processus et des conditions de production de la connaissance scientifique est une question toujours globalement ouverte, à laquelle des monuments de l'épistémologie comme la Critique de la Raison Pure de Kant, ou plus récemment l'étude par Kuhn de la "structure des révolutions scientifiques" [KUHN, 1970] ou le positionnement de Feyerabend pour une diversité des approches [FEYERABEND, 1993], ont apporté des éléments de réponse d'un point de vue philosophique. Un matériau plus empirique a été apporté également récemment avec les analyses quantitatives de la science, dans un sens une *épistémologie quantitative* qui va bien plus loin que des indicateurs bibliométriques purs [CRONIN et SUGIMOTO, 2014]. Les contributions s'intéressant à la complexité, c'est à dire étudiant des systèmes complexes en un sens très large, peuvent témoigner de la production de cadre de travail très divers qui peuvent être vus comme des éléments élémentaires de réponse à la question à un autre niveau ci-dessus. Nous utiliserons par la suite le terme *Cadre de Connaissances*, pour tout cadre tel ayant une composante épistémologique s'intéressant à la nature de la connaissance et à sa production. Pour illustrer, nous pouvons mentionner de tels cadres dans différents domaines, à différents niveaux, et avec des buts différents. Par exemple, [DURANTIN et al., 2017] explore les potentialités de coupler l'ingénierie avec des paradigmes du design to encourager l'innovation disruptive. Toujours en Gestion de Connaissances, utilisant la contrainte de l'innovation comme un avantage pour appréhender la nature complexe de la connaissance, [CARLILE, 2004] introduit les notions de frontières des domaines de connaissance et de processus de production. Introduisant également un framework meta, mais dans le champ de l'ingénierie des systèmes, [GEMINO et WAND, 2004] recommande l'utilisation

tion de grammaires pour comparer les techniques de Modélisation Conceptuelle. Les cadres de meta-modélisation peuvent aussi être compris comme des cadres de connaissance. [COTTINEAU et al., 2015a] décrit un cadre de multi-modélisation pour le test d'hypothèses dans la simulation des systèmes complexes socio-techniques. [GOLDEN, AGUIER et KROB, 2012] postule une formulation unifiée de la notion de système, ce qui inclut nécessairement différents types de connaissance sur un système correspondant à la description de ses différents composants.

Une explication possible pour une telle richesse est la nature fondamentalement réflexive de l'étude des Systèmes Complexes : à cause du choix plus grand pour la méthodologie et sur quels aspects du système mettre l'emphase, une partie significative d'une entreprise de modélisation ou de design est une exploration à un niveau meta. De plus, les études de la production de connaissance sont profondément ancrées dans la complexité, comme Hofstadter a bien souligné dans [HOFSTADTER, 1980] en rappelant l'existence de "boucles étranges", c'est à dire de boucles de rétroaction permettant la reflexivité comme une théorie s'appliquant à elle-même, dans ce qui constitue l'intelligence et l'esprit. L'intelligence artificielle est de fait un champ crucial au regard de nos réflexions, comme ses progrès impliquent une compréhension plus fine de la nature de la connaissance. [MOULIN-FRIER et al., 2017] introduit un meta-cadre pour une typologie générale des approches en intelligence artificielle, ce qui correspond à un cadre de connaissance non au sens propre mais dans un cas particulier d'application.

Le niveau des cadres présentés ci-dessus peut être très général mais reste conditionné à un certain champ ou discipline, et à une certaine approche ou méthodologie. Il n'existe à notre connaissance pas de cadre réalisant un exercice difficile, qui est de capturer une certaine structure de production de la connaissance à un niveau épistémologique, mais qui est conjointement pensée dans une perspective très appliquée, avec des conséquences directes pour la conception et la gestion de systèmes complexes. La contribution de cette partie propose de poser les bases pour un cadre réalisant cela dans le cas des Systèmes Complexes. Pour y parvenir, nous partons du postulat que la tension entre ces deux objectifs contradictoires est un atout pour éviter d'une part une généralité globale impossible et d'autre part une spécificité due à un domaine qui serait trop restrictive. En se basant sur l'idée des domaines de connaissance introduite par [LIVET et al., 2010], son aspect central est une approche cognitive de la science qui implique des processus de co-evolution entre les domaines de connaissance et leur supports. Une première ébauche de ce cadre a été présentée par [RAIMBAULT, 2017g], dans le cas particulier des systèmes complexes territoriaux comme étudiés par la géographie théorique et quantitative. Nous proposons de l'introduire ici par une dé-

marche inductive, c'est à dire en partant d'une étude de cas concrète qui a largement inspiré la construction du cadre, pour finir avec sa description générique.

9.3.1 Etude de cas

Genèse de la Théorie Evolutive Urbaine

La première étude de cas rappelle la construction de la *Théorie Evolutive Urbaine*¹³, une théorie géographique qui considère les systèmes territoriaux par une perspective complexe, développée depuis une vingtaine d'années environ. Nous étudions sa genèse par l'utilisation de méthodes mixtes, c'est à dire à la fois des interviews semi-dirigées avec des contributeurs principaux, et une analyse bibliométrique quantitative des publications principales. Les interviews ont été menées en suivant les standards méthodologiques classiques [LEGAVRE, 1996] pour assurer une interférence limitée des expériences de l'interviewer, mais sans le faire disparaître complètement afin de permettre un contexte précis favorable à la fluidité de l'interviewé. Nous utilisons ici des interviews¹⁴ avec Pr. D. Pumain qui a introduit et développé majoritairement la théorie, et Dr. R. Reuillon, dont la recherche sur le calcul intensif et distribué et l'exploration de modèles a été une pierre d'angle des développements les plus récents.

Pour commencer il est important de se rappeler un aperçu rapide du contenu de la théorie évolutive. Pour cela, consulter le deuxième pilier de notre théorie géographique en 9.1, qui en donne la substantifique moelle.

La caractéristique frappante dans cette construction est l'équilibre entre les différents *types* de connaissance, desquelles une typologie sera le point de départ de notre construction. La relation entre les considérations théoriques et les cas d'étude empiriques est fondamental. En effet l'article séminal [PUMAIN, 1997] est déjà positionné comme "un plaidoyer pour une théorie [...] moins ambitieuse, mais qui ne néglige pas les aller-retours avec l'observation". Nous pouvons maintenant nous tourner vers les entretiens pour mieux comprendre les implications de l'intrication des différents types de connaissance.

¹³ L'ambiguïté de l'adjectif *évolutive* fait gagner la théorie en subtilité, puisqu'il s'applique aussi bien au sens premier c'est à dire aux entités urbaines étudiées, mais aussi à un sens meta à la théorie elle-même, ce qui confirme un certain niveau de réflexivité de la théorie qui est essentiel comme développé en 3.3. Pour traduire le terme en anglais, il a été choisi "Evolutionary Urban Theory" par [PUMAIN et al., 2006], mais "Evolutive Urban Theory" convient aussi, mais il semble dans tous les cas difficile de transférer l'ambiguïté lors de la traduction.

¹⁴ Toutes les deux d'une durée environ une heure. Le son et les transcripts sont disponibles sous une Licence CC à <https://github.com/JusteRaimbault/Interviews> [RAIMBAULT, 2017d]. Les interviews sont en français et la traduction anglaise des passages cités dans l'article original est assurée par l'auteur.

D. Pumain retrace les idées germinales à son travail de maîtrise en 1968, quand “tout a commencé avec une question de données”. L’intérêt pour les villes, et pour le *changement dans les villes*, a été conduit par la disponibilité d’un jeu de données raffiné sur les flux migratoires à différentes dates. Egalement rapidement, est venue “la frustration des méthodes qui manquaient”, mais l’accès au centre de calcul (*outil technique*) a permis le test de méthodes et modèles nouvellement introduits, liés à l’approche de la complexité par Prigogine. Les méthodes restaient toutefois limitées pour capturer l’hétérogénéité des interactions spatiales. Un besoin progressivement spécifié et une rencontre fortuite, avec “une dame qui travaillait sur les réseaux de neurones et les modèles agents à la Sorbonne”, a conduit à une bifurcation et un nouveau niveau d’interaction entre modèles, théorie et connaissance empirique : en 1997, deux articles séminaux, l’un donnant la base théorique, l’autre introduisant le premier modèle Simpop, étaient publiés simultanément. A partir de ce point, il était clair que toute entreprise de modélisation était conditionnée à une connaissance empirique de cas d’étude géographiques et à des hypothèses théoriques à tester. Les méthodes et les outils techniques ont alors pris aussi un rôle nécessaire, avec des méthodes d’exploration de modèles spécifiques développées avec le logiciel OpenMole. R. Reuillon raconte qu’un saut qualitatif de connaissances a été rendu rapidement possible quand les méthodes d’exploration systématiques ont été introduites pour comprendre le comportement du modèle SimpopLocal. A la base, les géographes n’étaient pas sûr si le modèle fonctionnait seulement, dans le sens où il produisait les faits stylisés attendus comme l’émergence de la hiérarchie d’un système de villes. Des trajectoires satisfaisantes ont été trouvées par l’utilisation d’algorithmes génétiques de calibration, en calcul distribué sur grille [SCHMITT et al., 2014]. L’existence de multiples solution équivalentes pour les valeurs des paramètres est une barrière pour des question concrètes de nécessité ou suffisance d’un mécanisme donné du modèle agent. Ce besoin, venant du domaine de la connaissance empirique et théorique géographique, a mené à la conception d’un algorithme spécifique : le Calibration Profile, qui est une avancée méthodologique dans l’exploration de modèles [REUILLOU et al., 2015]. Ce cercle vertueux a été continué avec la famille de modèles Marius [COTTINEAU, 2014] et l’algorithme Parameter Space Exploration [CHÉREL, COTTINEAU et REUILLOU, 2015]. R. Reuillon évalue son impact du point de vue d’un informaticien : “Je ne suis pas sûr si les géographes étaient immédiatement conscients de la portée du résultat, c’était du lourd, les gens qui bossaient avec nous l’ont directement vu.” Cette vision positive est confirmée par D. Pumain, qui souligne les bénéfices de ces nouvelles méthodes pour la connaissance Géographique, et que c’était la première fois qu’une recherche menait à des publications à la frontière de la connaissance à la fois en géographie et en informatique.

En prenant du recul, émerge une typologie de domaines dans laquelle de la connaissance a été créée mais également nécessaire pour les autres domaines dans la genèse de la Théorie Evolutive Urbaine. La récolte des données et la construction de jeux de données est un premier pré-requis pour toute connaissance supplémentaire. A partir des données on extrait des faits stylisés empiriques, desquels sont déduits des hypothèses théoriques. La Théorie peut être testée pour falsification, dans le domaine empirique mais aussi par les modèles, par exemple par des expériences ciblées dans les modèles de simulation. De nouvelles méthodes sont alors développées pour mieux les explorer. Les outils sont cruciaux à chaque étape, pour implémenter un modèle, faire de la fouille de données ou collecter et formater les données par exemple. L'analyse précédente montre comment ces domaines sont interdépendants, et sont dans un sens *co-évolutifs*.

Nous supportons cette analyse qualitative par une analyse quantitative bibliométrique modeste. L'idée est d'étudier la structure du coeur du réseau de citations des publications principales construisant la Théorie Evolutive Urbaine. Nous construisons le réseau de citations comme décrit en Fig. 26, en utilisant l'outil de collection de données fournit par [RAIMBAULT, 2016d]¹⁵. Partant des deux publications séminales [PUMAIN, 1997] et [SANDERS et al., 1997], le réseau de citation inverse est obtenu à profondeur 2 (les références citant ces références initiales, et celles citant les citantes), en filtrant à la première étape sur les auteurs pour avoir au moins un des principaux contributeurs de la théorie (que nous prenons comme *Pumain, Sanders et Bretagnolle*, en accord avec l'entretien avec D. Pumain). Les noeuds de degré 1 sont supprimés, pour obtenir uniquement le coeur du réseau d'ego. On peut noter qu'il ne manque pas de lien entre les noeuds du premier niveau, puisque tous les liens citants ont été récupérés. Le réseau a une densité de 0.019, ce qui est plutôt élevé pour un réseau de citation, et la signature d'un haut niveau de dépendance entre les publications. En partant de deux noeuds distincts, nous aurions pu avoir en théorie des composantes connexes distinctes, mais comme attendu le réseau n'en a qu'une de par la nature fortement interconnectée des deux aspects. Pour analyser la structure de manière plus fine, nous détectons les communautés en utilisant l'algorithme de clustering de Louvain, et évaluons la modularité dirigée de la partition comme donnée par [NICOSIA et al., 2009]. Nous montrons en Fig. 26 une visualisation du réseau. Nous obtenons 7 communautés avec une valeur de modularité de 0.39. Pour assurer que cette valeur est significative, nous procédons à des simulations de Monte Carlo et distribuons de manière aléatoire les liens de citation 100 fois, en calculant à chaque fois la modularité des communautés dans le réseau aléatoire. Nous obtenons une modularité moyenne dirigée de $\bar{m} = 0.002 \pm 0.015$, rendant

¹⁵ L'ensemble du code et des données pour cette analyse sont disponibles à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo>

la modularité du réseau réel hautement significative (plus de 200 déviations standard). Nous analysons le contenu des communautés en examinant leur publications du premier niveau. Nous trouvons que les communautés sont globalement cohérentes avec les typologie des domaines : une pour les méthodes, trois sur la modélisation spatio-temporelle des systèmes urbains qui mélange empirique et modélisation, une conceptuelle, une sur les modèles Simpop, et une dernière sur les lois d'échelle qui est complètement empirique. Les *Data Papers* ne sont pas encore une pratique courante en géographie et des articles spécifiques au domaine des données ne peuvent être trouvés dans le réseau. Un taux de citation accru entre papiers du même domaine est dans tous les cas attendu à cause du standard scientifique de toujours situer une contribution au regard des travaux similaires. La valeur significative de la modularité confirme que les domaines sont cohérents au regard d'une certaine structure endogène de la production de connaissance.

Ingénierie

Après l'aperçu sur les domaines de connaissances extraits dans l'étude de cas précédente, nous proposons de prendre un point de vue similaire sur un exemple assez différent plus en relation avec la technologie et l'ingénierie. Nous interprétons ainsi des questions d'ingénierie liées au système de transport métropolitain parisien au travers du prisme des domaines de connaissance. En prenant l'exemple de l'automatisation progressive de la ligne 1, considérée largement comme une prouesse technique, de nombreuses études intégrant modélisation et études empiriques ont été conduite en préliminaire [BELMONTE et al., 2008]. L'utilisation et l'adaptation de méthodes particulières comme la modélisation basée-agent est cruciale pour le développement de transports autonomes innovants [BALBO, ADAM et MANDIAU, 2016]. Dans ce problème d'ingénierie, des solutions techniques comme les portes palières de quai peuvent être vues comme des outils qui évoluent également, et sont nécessaires pour qu'une nouvelle approche conceptuelle (*le transport automatique*) soit implémentée [FOOT, 2005]. Mais ils peuvent aussi interagir avec d'autres aspects de la connaissance conceptuelle, comme le management et l'organisation au sein de l'opérateur [FOOT, 1994]. L'aspect multi-dimensionnel complexe de l'innovation pour de tels systèmes avait déjà été souligné depuis longtemps comme le montre [HATCHUEL, PALLEZ et PÉNY, 1988]. D'autres aspects techniques, comme des problèmes d'ingénierie civile [MORENO REGAN, 2016], sont aussi mise en jeu pour développer une telle nouvelle approche, et ils nécessitent au moins les domaines empiriques et de modélisation, voire plus. Cet exemple relativement court illustre comment l'interprétation par domaines de connaissance peut être appliquée à l'ingénierie et au management de systèmes complexes industriels. Des détails spécifiques seraient né-

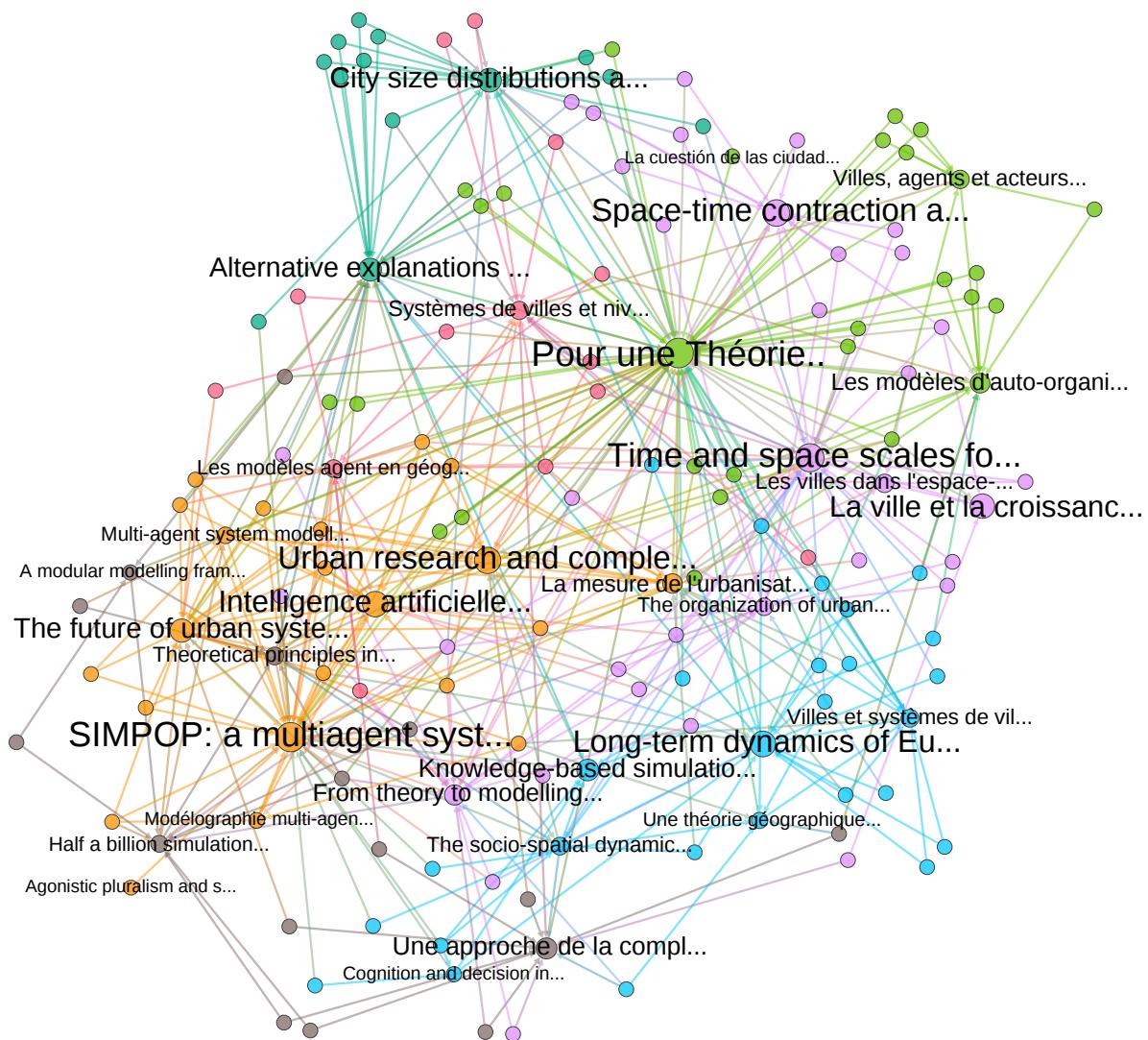


FIGURE 26 : Réseau de citations des publications principales de la Théorie Evolutive Urbaine. Le réseau est construit de la manière suivante : à partir des deux publications séminiales [PUMAIN, 1997] and [SANDERS et al., 1997], nous récupérons les publications les citant, filtrons sous la condition d'un des contributeurs principaux appartenant aux auteurs, récupérons encore les publications citantes et filtrons. Les noeuds sont les publications ($|V| = 155$), leur taille correspondant à la centralité de vecteur propre, et les liens sont les liens de citation dirigés ($|E| = 449$). La couleur donne les communautés obtenues par l'algorithme de clustering de Louvain (7 communautés, modularité 0.39).

cessaires pour une application plus en profondeur, mais nous proposons ici une preuve de concept.

9.3.2 Cadre de Connaissances

Nous pouvons à présent formuler le cadre de manière inductive. Comme déjà évoqué, il tire l'idée de domaines de connaissance en interaction du cadre introduit par [LIVET et al., 2010], mais étend ces domaines et prend une nouvelle position épistémologique, se concentrant sur les dynamiques co-évolutives entre agents et connaissances.

CONTRAINTE Pour être particulièrement adapté à l'étude et au management de la complexité, nous postulons que le cadre doit répondre à certaines contraintes, en particulier pour prendre en compte et même favoriser la *nature intégrative de la connaissance*, comme illustré par l'importance de l'interdisciplinarité et de la diversité dans les cas d'étude. Le cadre doit ainsi être favorable aux points suivants :

- Intégration des disciplines, puisque les Systèmes Complexes sont par essence à la croisée de champs multiples
- Intégration des domaines de connaissance, c'est à dire qu'aucun type particulier de connaissance ne doit être privilégié dans le processus de production¹⁶
- Intégration des types de méthodologie, en particulier dépasser les frontières artificielles entre méthodes "quantitatives" et "qualitatives", qui sont particulièrement fortes en sciences sociales et humanités classiques.

FONDATIONS ÉPISTÉMOLOGIQUES Le positionnement épistémologique du cadre est celui développé dans la première section de 3.3. Nous rappelons l'importance de la *perspective* [GIERE, 2010c], composée des agents, des objets représentés, du but et du medium (le modèle). L'approche par agents est fondamentale pour la cohérence du cadre.

DOMAINES DE CONNAISSANCE Nous postulons les domaines de connaissance suivants, avec leurs définitions :

- **Empirique.** Connaissance empirique d'objets du monde réel.
- **Théorique.** Connaissance conceptuelle plus générale, impliquant des constructions cognitives.
- **Modélisation.** Le modèle est le *medium* formalisé de la Perspective Scientifique, aussi divers que la classification de VARENNE des fonctions des modèles [VARENNE, 2010b] (voir ci-dessous).
- **Données.** Information brute qui a été collectée.

¹⁶ ce qui n'est pas incompatible avec des spécifications fonctionnelles très strictes, puisque des chemins divers sont possibles pour atteindre le même état final fixé

- **Méthodes.** Structures générées de production de connaissance.
- **Outils.** Proto-méthodes (implémentation des méthodes) et supports des autres domaines.

Nous prenons le parti de séparer Outils et Méthodes, pour insister sur le rôle de support des outils, et car le développement des deux est lié mais pas identique. De la même façon, le domaine des Données et le domaine Empirique sont distincts, car des nouveaux jeux de données n'impliquent pas systématiquement une nouvelle connaissance de faits empiriques, même si la construction des outils de captation de données souvent requiert une connaissance empirique. Le domaine de la Modélisation a un rôle central puisque nous postulons que *toute connaissance d'un système complexe nécessite un modèle*.

CO-ÉVOLUTION DES CONNAISSANCES Nous pouvons à présent formuler l'hypothèse centrale de notre cadre, qui est partiellement contenu dans le positionnement par rapport au Perspectivisme. Nous postulons que *toute construction de connaissance scientifique sur un système complexe¹⁷* est une perspective au sens de GIERE. Elle est composée de contenu de connaissance dans chacun des domaines, qui *co-évolue* entre eux et avec les autres éléments de la perspective, en particulier les agents cognitifs. La notion de co-évolution est prise au sens de [HOLLAND, 2012], c'est à dire d'entités étant fortement interdépendantes au sein de niches avec des relations causales circulaires et qui ont une certaine indépendance avec l'extérieur dans leur frontières. Nous notons l'importance de l'émergence faible au sens de BEDAU [BEDAU, 2002] dans la construction de la perspective à partir de la co-évolution de ses composants, comme il s'agit d'un niveau supérieur autonome qui peut être compris en lui-même, comme la connaissance scientifique peut être. Il faut aussi noter qu'une perspective n'a pas nécessairement des composants dans tous les domaines, mais devraient généralement en avoir dans la plupart.

APPLICATION Les types de modèles auquel notre cadre s'applique sont supposés être tous les modèles possibles en un sens très large,

¹⁷ Nous sommes convaincus que cet aspect intrigué de la production de connaissance est nécessairement présent pour les Systèmes Complexes, en écho à la remarque sur la réflexivité en introduction de la section. Même des *modèles simples* de systèmes complexes impliquent une complexité conceptuelle qui nécessite que la complexité de la connaissance soit présente pour être traduite. Cette dernière hypothèse pourrait liée à la nature de la complexité et la relation entre la complexité computationnelle et la complexité au sens de l'émergence faible, qui est suggérée par exemple par [BOLOTIN, 2014] qui explique l'émergence et la décohérence depuis le niveau quantique par la NP-complétude de la résolution des équations fondamentales. Ces considérations sont bien au delà de la portée de cette section (voir 3.3 pour une réflexion plus approfondie), et nous prenons comme une hypothèse que les systèmes complexes nécessitent de la connaissance complexe, tandis que de la connaissance simple (au sens de domaines et agents non co-évolutifs) *peut* exister pour des systèmes simples.

puisque GIÈRE désigne par modèle tout *medium* d'une perspective. Une vue fonctionnelle des modèles comme VARENNE introduit [VARENNE, 2010b] (introduisant une typologie des modèles par leur fonctions, par exemple les modèles explicatifs, les modèles de simulation, les modèles prédictifs, les modèles de compréhension, les modèles interactifs, etc.) est un moyen d'appréhender leur variété. Il est aussi possible de le voir en terme de classifications plus classiques, et l'appliquer au modèles mathématiques, statistiques, de simulation, de données, ou conceptuels par exemple. Concernant les contraintes données précédemment, comme toutes les connaissances sont en coévolution, aucun domaine n'est privilégié en particulier. Aucune discipline non plus, puisque celles-ci auront leur différents aspects contenus dans les domaines, et finalement les méthodes qualitatives et quantitatives seront présentes et nécessaire dans la majorité. Nous montrons en Fig. 27 une projection des domaines de connaissance comme un réseau complet, pour illustrer de quoi peuvent être composées les relations entre domaines.

9.3.3 Discussion

Portée d'application

Nous insistons sur le fait que notre cadre ne prétend pas introduire une épistémologie générale de la connaissance scientifique, mais loin de cela est plutôt ciblé vers une réflexivité dans la compréhension des systèmes complexes. Le niveau de généralité est à niveau très différent, mais le but d'implications pratiques dans la compréhension de la complexité contribue à un certain caractère générique dans les applications. Il est de plus particulièrement adapté à l'étude des Systèmes Complexes, puisque des approches plus réductionnistes peuvent gérer des productions de connaissance plus compartimentées, tandis que l'intégration des disciplines et des échelles et donc des domaines de connaissance a été souligné comme crucial pour l'étude de la complexité.

Vers une Formalisation

Le cadre de connaissances reste à un niveau épistémologique, et son application pourrait être formalisée de manière plus systématique. Pour cela, il faudrait reprendre partiellement le cadre développé dans la section précédente 9.2. Rappelons les éléments clés et comment ceux-ci peuvent s'articuler. L'aspect principal est le couplage d'une formalisation du modèle du système avec celle de la perspective. Une perspective serait définie comme une *Dataflow Machine M* au sens de [GOLDEN, AIGUER et KROB, 2012] qui donne un moyen pratique pour la représenter et pour introduire les échelles de temps et les données, à laquelle est associée une ontologie O au sens de [LIVET

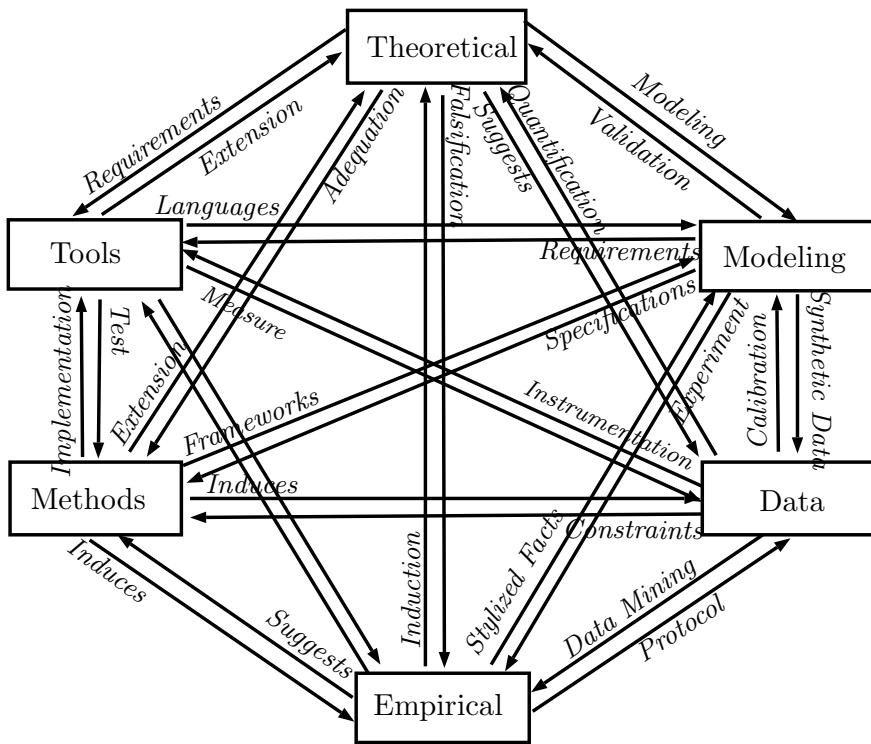


FIGURE 27 : Projection d'une perspective comme un réseau complet des domaines de connaissance. Pour illustrer les domaines et les processus d'interaction possibles entre ceux-ci, nous faisons l'exercice d'essayer de qualifier toutes les relations binaires possibles entre les domaines. Cela ne reflète en rien la structure réelle du cadre, mais est une aide pour considérer ce que les interactions peuvent être. Il faut noter que la nature des relations n'est pas toujours la même ici, certaines étant des contraintes, d'autres des transferts de connaissance, d'autres processus à l'intérieur d'autres domaines comme les données synthétiques qui est une méthodologie. Cela montre que certains domaines agissent comme catalyseurs pour les relations entre les autres, dans cette configuration de réseau, ce qui correspond en fait à une situation de co-évolution.

et al., 2010], i.e. un ensemble d'éléments dont chacun correspond à une entité (qui peut être un objet, un agent, un processus, etc.) du monde réel. Le motif et l'agent porteur de la perspective sont contenus dans l'ontologie s'ils font sens pour étudier le système. Décomposer l'ontologie en éléments atomiques $O = (O_j)_j$ et introduire une relation d'ordre entre les éléments des ontologies basée sur l'émergence faible ($O_j \succcurlyeq O_i$ si et seulement si O_j émerge faiblement de O_i) devrait fournir une décomposition canonique de la perspective contenant la structure du système. Le défi serait ensuite de lier cette décomposition avec la décomposition canonique de la *Dataflow Machine* postulée par [GOLDEN, AIGUER et KROB, 2012], et ensuite définir les

domaines de connaissance au sein de ce couplage : les données sont dans les flots des machines, le modèle est la machine, l'empirique et le théorique dans les ontologies, les méthodes dans la structure de l'arbre. Une telle entreprise avec des opérations cohérentes entre les éléments est cependant hors de notre portée pour l'instant, mais serait un développement puissant.

Nous avons étudié par des méthodes mixtes la construction d'une théorie scientifique en géographie théorique et quantitative, et à partir de cela introduit de manière inductive un cadre de connaissances visant comprendre la production de connaissances sur un système complexe comme un système complexe elle-même, plus précisément une perspective avec des composantes co-évolutives au sein de domaines de connaissances interdépendants. On peut noter que cette approche est totalement réflexive puisque plusieurs de ces composantes ont été nécessaires. Nous postulons que ce cadre peut être un outil utile pour étudier la complexité et gérer des systèmes complexes, puisqu'il explicite certains choix et directions de développements qui pourraient autrement être inconscients.

Co-construction des théories et modèles en géographie quantitative : une synthèse de nos contributions

Nous concluons ce chapitre d'ouverture par une mise en perspective cohérente des diverses contributions de la thèse, du point de vue de l'illustration de la co-évolution des connaissances dans différents domaines, et de boucler la boucle par un retour sur la construction de la théorie géographique. Comme précisé en préambule, un mode de lecture linéaire serait trop réducteur, puisque la plupart des travaux s'enrichissent mutuellement quel que soit leur domaine et leur portée, et un compte-rendu linéaire, au delà d'être intrinsèquement appauvrissant, est en quelque sorte un mensonge par omission de l'ensemble des interactions complexes entre les pans de connaissance produite. Bien sûr l'exercice de synthèse et la capacité à faire rentrer dans un cadre formaté imposé, sont louables, voir souhaitables dans l'état actuel des conditions de production scientifiques. Mais une posture fondamentale que nous prendrons et défendrons tout au long de ce travail est celle d'une science anarchiste proposée par FEYERABEND, qui sans être prise purement littéralement et mise en contexte, est extrêmement fructifiante pour proposer des changements de paradigmes et s'émanciper de travaux *mainstream* dont les bases et la légitimité semblent s'enrichir malgré les critiques croissantes. L'écriture d'une monographie extrêmement formatée ne présente généralement que peu d'intérêt de par le caractère contraint de l'exercice (combien d'interminables chapitres "état de l'art" et "problématique" ou "enjeux sociaux" témoignent d'une platitude au point de vouloir arrêter la lecture d'un ouvrage par ailleurs remarquable, ce qui s'est sûrement passé dans notre cas d'ailleurs), et paraît relativement vaine vu

la destinée de prendre la poussière dans une étagère obscure d'un laboratoire obscur, sans être sauvé par la mise en ligne vu la langue imposée¹⁸. On se rêve d'imaginer une thèse entièrement digitale et dont le cheminement du lecteur tracé dans le support numérique serait à l'origine d'une multitude de visions possibles, traduisant effectivement la complexité du processus de construction, et des perspectives d'enrichissement innombrables par une rétroaction et une interaction avec les lecteurs, c'est à dire sortir du mode de présentation linéaire, comme déjà soutenu en introduction. L'invention de nouveaux modes de communication scientifiques est un défi urgent à part entière, et notre ébauche de réflexivité développée en Appendice F cherche à y contribuer.

La construction de théories géographiques, dans le cadre d'une Géographie Théorique et Quantitative, s'effectue par itérations dans une dynamique de co-évolution avec les efforts empiriques et de modélisation [LIVET et al., 2010]. Parmi les nombreux exemples, on peut citer la théorie évolutive des villes (co-construite par un spectre de travaux s'étendant par exemple des premières propositions de [PUMAIN, 1997] jusqu'aux résultats matures présentés dans [PUMAIN, 2012a]), l'étude du caractère fractal des structures urbaines (par exemple de [FRANKHAUSER, 1998] à [FRANKHAUSER, 2008]) ou plus récemment le projet Transmondyn visant à enrichir la notion de transition des systèmes de peuplement (ouvrage à paraître). Cette communication propose un format original en s'inscrivant dans cette lignée, par la synthèse de différents travaux empiriques et de modélisation menés conjointement avec l'élaboration d'appareils théoriques visant à mieux comprendre les relations entre territoires et réseaux de transports. L'originalité de cette contribution réside à la fois dans la synthèse de travaux très divers pourtant reliés en filigrane, et dans la proposition d'une théorie géographique spécifique s'appuyant sur cette synthèse en seconde partie.

POURQUOI UNE THÉORIE ET DES MODÈLES DE CO-ÉVOLUTION
Notre première entrée prend un point de vue d'épistémologie quantitative pour tenter d'expliquer le fait que, si la co-évolution entre territoires et réseaux a par exemple été prouvée par [BRETAGNOLLE, 2009a], la littérature est très pauvre en modèles de simulation endogenéisant cette co-évolution. Une exploration algorithmique de la littérature a été menée dans [RAIMBAULT, 2015a], suggérant un cloisonnement des domaines scientifiques s'intéressant à ce sujet. Des méthodes plus éla-

¹⁸ Ce qui relève bien sûr par ailleurs d'une problématique bien plus complexe que la simple audience [TARDY, 2004] et la richesse des pensées scientifiques permises par l'utilisation de différentes langues n'est pas discutable ainsi que la légitimité d'organisations comme l'ASRDLF. Mais c'est bien cette audience qui nous pose problème ici et dans ce cas il est quasiment aussi vieux jeu pour une école doctorale d'imposer le français comme langue d'écriture que le discours du consul et son snobisme d'énarque rapportés en 1.2.

borées ainsi que les outils correspondants (collecte et analyse des données), couplant une analyse sémantique au réseau de citations, ont été développées pour renforcer ces conclusions préliminaires [RAIMBAULT, 2016d], et les premiers résultats au second ordre semblent confirmer l'hypothèse d'un domaine peu défriché car à l'intersection de champs ne dialoguant pas nécessairement aisément. Ces premiers résultats d'épistémologie quantitative confirment l'intérêt d'une modélisation couplant des processus relevant de différentes échelles et domaines d'études, et surtout l'intérêt de l'élaboration d'une théorie propre.

ETUDES EMPIRIQUES Le premier axe pour les développements en eux-mêmes consiste en des analyses empiriques. Une étude des corrélations spatiales statiques entre mesures de forme urbaine (indicateurs morphologiques calculés sur la grille de population eurostat) et mesures de forme de réseau (topologie du réseau routier issu d'OpenStreetMap), sur l'ensemble de l'Europe à différentes échelles, a pu révéler la non-stationnarité et la multi-scalarité spatiale de leurs interactions [RAIMBAULT, 2016c]. Cet aspect a aussi été mis en évidence dans l'espace et le temps à une échelle microscopique lors de l'étude des dynamiques d'un système de transport [RAIMBAULT, 2016e], conjointement avec l'hétérogénéité des processus pour un autre type de système [RAIMBAULT, 2015b]. Ces faits stylisés valident pour l'instant l'utilisation de modèles de simulation complexes, pour lesquels des premiers efforts de modélisation ont ouvert la voie vers des modèles plus élaborés.

MODÉLISATION A l'échelle mesoscopique, des processus d'agrégation-diffusion ont été prouvés suffisant pour reproduire un grand nombre de formes urbaines avec un faible nombre de paramètres, calibrés sur l'ensemble du spectre des valeurs réelles des indicateurs de forme urbaine pour l'Europe. Ce modèle simple a pu, à l'occasion d'un exercice méthodologique explorant le possibilité de contrôle au second ordre de la structure de données synthétiques [RAIMBAULT, 2016a], être couplé faiblement à un modèle de génération de réseau, démontrant une grande latitude de configurations potentiellement générées. L'exploration de différentes heuristiques autonomes de génération de réseau a par ailleurs été entamée [RAIMBAULT et GONZALEZ, May 2015], pour comparer par exemple des modèles de croissance de réseau routier basés sur l'optimisation locale à des modèles inspirés des réseaux biologiques : chacun présente une très grande variété de topologies générées. A l'échelle macroscopique, un modèle simple de croissance urbaine calibré dynamiquement sur les villes françaises de 1830 à 2000 (base Pumain-Ined) a permis de démontrer l'existence d'un effet réseau de par l'augmentation de pouvoir explicatif du modèle lors de l'ajout d'un effet des flux transitant par

un réseau physique, tout en corrigeant le gain dû à l'ajout de paramètres par la construction d'un Critère d'Information d'Akaike empirique [RAIMBAULT, 2016f]. Cet ensemble de modèles se positionne avec un objectif de parcimonie et dans une perspective d'application en multi-modélisation. Dans une démarche basée-agent plus descriptive et donc par un modèle plus complexe, [LE NÉCHET et RAIMBAULT, 2015] décrit un modèle de co-évolution à l'échelle métropolitaine (modèle Lutecia) qui inclut en particulier des processus de gouvernance pour le développement des infrastructures de transport. Même si ce dernier modèle est toujours en exploration, les premières études de la dynamique montre l'importance du caractère multi-niveau du développement du réseau de transport pour obtenir des motifs complexes de réseaux et de collaboration entre agents. L'ensemble de ces premiers efforts de modélisation, bien qu'ils ne soient pas majoritairement centrés sur des modèles de co-évolution à proprement parler, supportent les premiers fondements théoriques que nous proposons par la suite.

CONSTRUCTION D'UNE THÉORIE GÉOGRAPHIQUE Nous revoyons enfin sous l'oeil de la co-evolution des domaines la théories construite en 9.1. Nous insistons ici sur son caractère intégratif permettant de joindre Théorie Evolutive et Morphogenèse. En se basant sur les travaux précédents, nous proposons de joindre deux entrées pour la construction d'une théorie géographique ayant un focus privilégié sur les interactions entre territoires et réseaux. La première est par la notion de *morphogénèse*, qui a été explorée d'un point de vue interdisciplinaire dans [ANTELOPE et al., 2016]. Pour notre part, la morphogenèse consiste en l'émergence de la forme et de la fonction, via des processus locaux autonomes dans un système qui exhibe alors une architecture auto-organisée. La présence d'une fonction et donc d'une architecture distingue les systèmes morphogénétiques de systèmes simplement auto-organisés (voir [DOURSAT, SAYAMA et MICHEL, 2012]). De plus, les notions d'autonomie et de localité s'appliquent bien à des systèmes territoriaux, pour lesquels on essaye d'isoler les sous-systèmes et les échelles pertinentes. Les travaux sur la génération de forme urbaine calibrée par des processus autonomes, les premiers travaux sur la génération de réseaux par de multiples processus également autonomes, et des travaux plus anciens étudiant un modèle simple de morphogenèse urbaine qui suffisait à reproduire des motifs de forme stylisés [RAIMBAULT, BANOS et DOURSAT, 2014], nous suggèrent la possible existence de tels processus au sein des systèmes territoriaux. D'autre part, le cadre d'un théorie évolutive des villes est plébiscité par nos résultats empiriques, qui montrent le caractère non-stationnaire, hétérogène, multi-scalaire des systèmes urbains. Pour rester le plus général possible, et comme nos résultats à la fois empiriques et de modélisation (génération de formes quelconques par

le modèle d'agrégation-diffusion par exemple) s'appliquent aux systèmes territoriaux en général, nous nous plaçons dans le cadres de territoires humains de Raffestin [RAFFESTIN, 1988], c'est à dire "la conjonction d'un processus territorial avec un processus informationnel", qui peut être interprété dans notre cas comme le système complexe socio-techno-environnemental que constitue un territoire et les agents et artefacts qui y interagissent. L'importance des réseaux est soulignée par nos résultats sur la nécessité du réseau dans le modèle de croissance macroscopique : nous proposons alors de parler de *Systèmes Territoriaux Complexes en Réseaux*, en ajoutant au plongement du territoire dans la théorie évolutive la particularité qu'il existe des composantes cruciales qui sont les réseaux (de transport en l'occurrence), dont l'origine peut être expliquée par la théorie territoriale des réseaux de Dupuy [DUPUY, 1987]. Nous spéculons alors l'hypothèse suivante afin de réconcilier nos deux approches : **l'existence de processus morphogénétiques dans lesquels les réseaux ont un rôle crucial est équivalente à la présence de sous-systèmes dans les systèmes territoriaux complexes en réseaux, qu'on définit alors comme co-évolutifs.** Cette proposition a de multiples implications, mais a typiquement guidé notamment les choix de modélisation vers une méthodologie modulaire et de multi-modélisation afin d'essayer d'exhiber des processus morphogénétiques, ainsi que les travaux empiriques vers une étude plus poussée des correlations, causalités (dans le cas de séries temporelles) et recherche de décompositions modulaires des systèmes.

* * *

*

CONCLUSION DU CHAPITRE

Dans une logique de lecture linéaire, cette ouverture par l'introduction de cadres théoriques selon divers points de vue, devrait avoir synthétisé et rassuré sur les questions ouvertes a priori réglées dans leur majorité - seul la conclusion pouvant encore apporter une chute dans la narration. Il s'agit d'un malentendu, et le lecteur qui voudrait être rassuré aurait du s'arrêter au Chapitre précédent, à la fin duquel nous avions fait un tour relativement conséquent des approches proposées. Ce chapitre ouvre en fait un gouffre, et fait prendre conscience que la portée des connaissances est extrêmement embryonnaire. Pour donner une allégorie, nous serions un peu dans la situation du périphérie de Mercure et du spectre de l'atome qui étaient des détails négligeables pour la physique classique à la fin du 19ème siècle, et ont mené aux gigantesques développements au cours du 20ème que sont la physique quantique et la relativité générale. Les questions soulevées par chacun des niveaux sont fondamentales pour l'étude des systèmes territoriaux complexes mais aussi des systèmes complexes en général. La théorie proposée en 9.1 pointe à nouveau la question de la non-stationnarité spatio-temporelle et la non-ergodicité dans un contexte multi-échelle, que nous postulons cruciale mais très peu comprise. On distingue aussi la difficulté d'intégration de théories existantes ce qui implique une compréhension le couplage de modèle. Ce problème est au coeur du cadre formel développé par la suite 9.2, qui soulève aussi des questions d'imbrication d'échelles. Le problème d'obtenir une structure algébrique cohérente avec une action de monoïde sur les données implique une intégration de la théorie de KROB, ce qui questionne plus généralement l'intégration des approches d'ingénierie système (systèmes complexes "industriel") avec celles de systèmes complexes naturels. La possibilité de théorie intégratives est soulevée par l'introduction du cadre de connaissance 9.3, qui pose également des problèmes plus généraux de production des connaissances et de nature de la complexité que nous avions brièvement abordé d'un point de vue épistémologique en 3.3. Nous proposons de synthétiser une partie de ces diverses question ouvertes dans un projet de recherche cohérent sur un long terme mais incluant des premières pistes concrètes immédiates, que nous présenterons en ouverture.

★ ★

★

CONCLUSION

A building is never used the way it was designed, that is a reality which grasping makes the difference between good and excellent architects. The effective functional use give sense to any construction. So goes it for a knowledge edifice. We shall now take a look back on what we constructed and try to take a step back. This part develops first theoretical apparels emerging from the various aspects already tackled. It then proposes to extract fundamental open questions that future research on territorial complex systems will have to tackle in the incoming decades.

OUVERTURES

PERSPECTIVES THÉMATIQUES ET GÉNÉRALES

Développement Spécifiques

Le mode de communication scientifique actuel est loin d'être optimal et les initiatives se multiplient pour proposer des modèles alternatifs : la revue post-publication en est une, l'utilisation de systèmes de contrôle de version et de dépôts publics une autre, ou la publication éclair de pistes de recherche (*Journal of Brief Ideas*). Les descriptions courtes de pistes de recherche sont souvent reléguées à la discussion ou la conclusion des articles, qui s'écrivent de manière conventionnelle, souvent avec un biais pour justifier *a posteriori* l'intérêt de *sa nouvelle méthode* qu'il faut malheureusement vendre. On fait alors des plans sur la comète, propose des développements ayant peu de rapport, ou des domaines d'application *qui auront un impact* (lire qui sont à la mode ou qui reçoivent le plus de financements à la période de l'écriture). Ce manuscrit tombe bien évidemment partiellement sous ces critiques, et encore plus les articles qui lui sont associés.

Nous proposons dans cette section un exercice pas forcément conventionnel : proposer des idées et développements possibles, en s'efforçant de concrétiser les questions de recherche et/ou points techniques autant que possible, afin que ceux-ci ne s'apparentent pas à une bouteille à la mer.

*Epistémologie Quantitative**Modèles Multi-scalaires**Vers des Modèles Opérationnels*

VERS UN PROGRAMME DE RECHERCHE

Pour une Géographie Intégrée Alternative

Comme déjà souligné en citant REY, les bouleversements techniques et méthodologiques qu'une discipline peut subir sont souvent accompagnés de profondes mutations épistémologiques, voire de la nature même de la discipline. Il est impossible de juger si l'état actuel des connaissances est transitoire, et s'il l'est quelle est le régime stable qui terminerait la transition s'il en existe un. La spéulation est le seul moyen de lever partiellement le voile, sachant que celle-ci sera nécessairement auto-réalisatrice : proposer des visions ou des programmes de recherche oriente les moyens et questions. L'incomplétude théorique en physique, lorsqu'il s'agit par exemple de lier relativité générale et physique quantique, c'est à dire le microscopique stochastique au macroscopique déterministe, orientent les visions du futur de la discipline qui elle-même conditionnent les actions concrètes qui dans ce domaine sont indispensables (financement du CERN ou de l'interféromètre d'ondes gravitationnelles spatial LISA). En géographie, même si les investissements techniques sont incomparables, ceux-ci existent (accès aux moyens de calcul, financement de laboratoires intégrés, etc.) et sont déterminés également par les perspectives pour la discipline. Nous proposons ici une vision et un manifeste d'une nouvelle géographie, qui est déjà en train de se faire et dont les bases sont solidement construites petit à petit. L'aventure de l'ERC Geodivercity en est l'allégorie, d'autant plus qu'elle a confirmé la plupart des directions professées par BANOS [BANOS, 2017]. L'intégration de la théorie, de l'empirique, de la modélisation, mais aussi de la technique et de la méthode, n'a jamais été aussi creusée et renforcée que dans les divers développements du projet. Sans l'accès à la grille de calcul et aux nouveaux algorithmes d'exploration permis par OpenMole, les connaissances tirées du modèle SimpopLocal auraient été moindres, mais les développements techniques ont aussi été conduits par la demande thématique.

Nous proposons un cadre de connaissances pour les études ayant une composante quantitative, ou plus précisément se posant dans la lignée de la Géographie Théorique et Quantitative (TQG). Ce cadre tente de répondre aux contraintes suivantes : (i) transcender les frontières artificielles entre quantitatif et qualitatif ; (ii) ne pas favoriser de composante particulière parmi les moyens de production de connaissance (aussi divers que l'ensemble des méthodes qualitatives et quantitatives classiques, les méthodes de modélisation, les approches théoriques, les données, les outils), mais bien le développement conjoint de chaque composante. Nous étendons le cadre de connaissances de [livet2010ontology], qui consacre le triptyque des domaines empiriques, conceptuels et de la modélisation, en y ajoutant les domaines à

part entière que sont les méthodes, les outils (qu'on peut voir comme des proto-méthodes) et les données. Les interactions entre chaque domaine sont détaillées, comme par exemple le passage des méthodes vers les outils qui consiste en l'implémentation, ou le passage de l'empirique aux méthodes comme prospection méthodologique. Toute démarche de production de connaissance, vue comme une *perspective* au sens de [GIERE, 2010c], est une combinaison complexe des six domaines, les fronts de connaissance dans chacun étant en co-évolution. Nous nommons notre cadre de connaissance *Géographie Intégrée*, pour souligner à la fois l'intégration des différents domaines mais aussi des connaissances qualitatives et quantitatives, puisque les deux se fondent dans chacun des domaines.

Axes de Recherche

TODO : idée : lister les principaux contributeurs etc.; quoi est compatible avec quoi quest ce quon pourrait coupler etc; faire analyse epistemo quanti.

TODO : add somewhere something on the link “more systematic evidence-based”-politics in science - less dogmatism. or what place for evidence-based research in social science? linked with quanti-quali : BEYOND classical separations, evidence-based and complex systems allow integration, socially responsible, but evidence-based and systematic..

TODO : one axis or opening on different types of complexities : “Complexity, Complexities, and Complex Knowledges”. → already in positioning

C (JR) : evoquer ouverture des cours, formation interdisciplinaire etc. : pas ici, plutôt en ouverture finale ?

NON-STATIONNARITÉ, NON-ERGODICITÉ ET DÉPENDANCE AU CHEMIN

COUPLAGE DES MODÈLES ET APPROCHES **TODO : different approaches to coupling / coupling to a certain degree using Kolmogorov etc : specific section or insert here ?**

CONSTRUIRE DES OUTILS DE VALIDATION POUR LES MODÈLES DE SIMULATION

EPISTÉMOLOGIE QUANTITATIVE ET EXPÉRIMENTALE POUR UNE INTÉGRATION EFFECTIVE Le mantra du mariage entre qualitatif et quantitatif est asséné mécaniquement par de nombreux auteurs, mais lorsqu'il s'agit de mise en application, on peut se permettre de soupçonner dans le meilleur des cas une naïveté, dans le pire des cas une hypocrisie. Quel sens à faire semblant de faire des analyses

quantitatives en tartinant des pages de régression linéaires dont le R^2 ne dépasse pas 0.1 ? Quel sens à faire semblant de détenir une connaissance qualitative fine pour justifier la mise en place de modèles relevant de l'usine à gaz technocratique ?¹⁹

POUR UNE SCIENCE TOTALEMENT OUVERTE **TODO : brosser ici directions vers lesquelles travailler; intégrer faits dans positionnements**

La transparence et mise en disponibilité des données brutes ou au moins pré-traitées, et du code informatique produisant les sorties de simulation ou les figures, semble être plutôt l'exception que la règle en géographie. Comme l'assène BANOS qui y dédie un de ses commandements, "le modélisateur n'est pas le gardien de la vérité prouvée", et comme rappelé en chapitre B, une reproductibilité parfaite des résultats est nécessaire pour une reconnaissance d'une quelconque valeur par la communauté scientifique, comme une théorie qui ne fournit pas de possibilité de falsification ne peut être considérée comme scientifique comme l'a introduit POPPER. Des expériences de revue pour *Cybergeo* ont confirmé à l'unanimité ce problème fondamental. Rappelons que la revue *PNAS* exige les données brutes et tableau produisant toute figure, pour prévenir tout biais de visualisation qu'il soit volontaire (ce qui est rédhibitoire et conduit à un signalement) ou non.

Les observateurs soulevant le caractère détraqué du mode actuel de publication scientifique sont nombreux. Un papier n'est pas un format compréhensible ni vraiment reproductible, et pousse au biais. Comme me le rappelait un ami qui s'est spécialisé de manière admirable dans l'acceptation de papiers extrêmement techniques par des *top-journals* économiques, écrire de façon à être accepté est "un jeu" dont les règles sont subtiles et qu'il faut maîtriser pour faire carrière. Selon notre positionnement, un tel mode de communication est contraire à l'honnêteté et l'intégrité intellectuelle nécessaires à une science éthique et ouverte. De la même façon que nous soutenons qu'une présentation linéaire d'un travail de thèse est trop fortement réducteur

¹⁹ cette remarque est partiellement une auto-critique, puisqu'il faut rappeler le caractère très peu qualitatif de notre travail

CONCLUSION

*Explorer sans relâche les systèmes géographiques...
- ARNAUD BANOS*

Le lecteur qui aura tenu jusqu'ici et qui a la mémoire solide ou bien sélective, ou encore qui aura adopté un style de lecture roman policier, se plaindra du manque d'originalité dans l'origine des citations introducives. Ce n'est pas anodin si les positions de BANOS, simples mais efficaces et profondes, ouvrent et ferment ce travail : les "9 principes de Banos" sont implicitement présents dans la majorité des travaux menés et perspectives ouvertes.

C (JR) : Le démon de Banos : est capable de faire de l'interdisciplinaire et du disciplinaire sans se perdre, et respecte les 9 points.

BIBLIOGRAPHIE

- ABADIE, Alberto et al. (2010). "Synthetic control methods for comparative case studies : Estimating the effect of California's tobacco control program". In : *Journal of the American Statistical Association* 105.490.
- ABBAS, Assad et al. (2014). "A literature review on the state-of-the-art in patent analysis". In : *World Patent Information* 37, p. 3–13.
- ACEMOGLU Daron, Akcigit Ufuk et William KERR (2016). "Innovation Network". In : *Proceedings of the National Academy of Sciences (forthcoming)*.
- ACHIBET, Merwan et al. (2014). "A Model of Road Network and Buildings Extension Co-evolution". In : *Procedia Computer Science* 32, p. 828–833.
- ADAMATZKY, Andrew et Jeff JONES (2010). "Road planning with slime mould : if Physarum built motorways it would route M6/M74 through Newcastle". In : *International Journal of Bifurcation and Chaos* 20.10, p. 3065–3084.
- ADAMS, Stephen (2010). "The text, the full text and nothing but the text : Part 1 - Standards for creating textual information in patent documents and general search implications". In : *World Patent Information* 32.1, p. 22–29. URL : <https://ideas.repec.org/a/eee/worpat/v32y2010i1p22-29.html>.
- AGHION, Philippe et Peter HOWITT (1992). "A Model of Growth through Creative Destruction". In : *Econometrica* 60.2, p. 323–51. URL : <https://ideas.repec.org/a/ecm/emetrp/v60y1992i2p323-51.html>.
- AGHION, Philippe et al. (1998). *Endogenous growth theory*. MIT press.
- AGHION, Philippe et al. (2015). *Innovation and Top Income Inequality*.
- AKCIGIT, Ufuk et al. (2013). *The Mechanics of Endogenous Innovation and Growth : Evidence from Historical US Patents*. Rapp. tech. Citeseer.
- ALBA, Martha de et D Miguel Ángel AGUILAR (2012). "Déplacements urbains et interaction sociale : le cas du système de transport collectif par métro dans la ville de Mexico". In : *Bulletin de psychologie* 1, p. 19–32.
- ALI, A. et al. (June 2014). *Les Eco-quartiers lus par la mobilité : vers une évaluation intégrée*. Rapp. tech. Ecole des Ponts ParisTech.
- ALLEN, P. et M. SANGLIER (1979). "A dynamic model of growth in a central place system". In : *Geographical Analysis* 11, p. 256–272.
- ANAS, Alex et al. (1998). "Urban Spatial Structure". English. In : *Journal of Economic Literature* 36.3, pp. 1426–1464. ISSN : 00220515. URL : <http://www.jstor.org/stable/2564805>.
- ANDERSON, Philip W (1972). "More is different". In : *Science* 177.4047, p. 393–396.

- ANGRIST, Joshua D et al. (1996). "Identification of causal effects using instrumental variables". In : *Journal of the American statistical Association* 91.434, p. 444–455.
- ANTELOPE, Chenling et al. (2016). "An Interdisciplinary Approach to Morphogenesis". In : *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016*.
- ARCAUTE, E. et al. (2013). "Constructing cities, deconstructing scaling laws". In : *ArXiv e-prints*. arXiv : 1301.1674 [physics.soc-ph].
- ARCHIBUGI, Daniele et Mario PIANTA (1992). "Specialization and size of technological activities in industrial countries : The analysis of patent data". In : *Research Policy* 21.1, p. 79 –93. ISSN : 0048-7333. DOI : [http://dx.doi.org/10.1016/0048-7333\(92\)90028-3](http://dx.doi.org/10.1016/0048-7333(92)90028-3). URL : <http://www.sciencedirect.com/science/article/pii/004873392900283>.
- ARTHUR, W. Brian (2015). *Complexity and the Shift in Modern Science*. Conference on Complex Systems, Tempe, Arizona.
- AUDRETSCH, David B et Maryann P FELDMAN (1996). "R&D spillovers and the geography of innovation and production". In : *The American economic review* 86.3, p. 630–640.
- AXTELL, Robert L (2016). "120 million agents self-organize into 6 million firms : a model of the US private sector". In : *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents et Multiagent Systems, p. 806–816.
- BADARIOTTI, Dominique et al. (2007). "Conception d'un automate cellulaire non stationnaire à base de graphe pour modéliser la structure spatiale urbaine : le modèle Remus". In : *Cybergeo : European Journal of Geography*.
- BAFFI, Solène (2016). "Railways and city in territorialization processes in South Africa : from separation to integration?" Theses. Université Paris 1 - Panthéon Sorbonne. URL : <https://halshs.archives-ouvertes.fr/tel-01389347>.
- BAIS, Sander et al. (2010). *In praise of science : curiosity, understanding, and progress*. MIT Press.
- BALBO, Flavien et al. (2016). "Positionnement des systèmes multi-agents pour les systèmes de transport intelligents". In : *Revue des Sciences et Technologies de l'Information-Série RIA : Revue d'Intelligence Artificielle* 30.3, p. 299–327.
- BANOS, Arnaud (2001). "A propos de l'analyse spatiale exploratoire des données". In : *Cybergeo : European Journal of Geography*.
- (2013). "Pour des pratiques de modélisation et de simulation libérées en Géographies et SHS". In : *HDR. Université Paris 1*.
- (2017). "Knowledge Accelerator' in Geography and Social Sciences : Further and Faster, but Also Deeper and Wider". In : sous la dir. de Denise PUMAIN et Romain REUILLOU. in *Urban Dynamics et Simulation Models*. Springer.

- (Décembre 2013). "Pour des pratiques de modélisation et de simulation libérées en Géographie et SHS". In : *Thèse d'Habilitation à Diriger des Recherches, UMR CNRS 8504 Géographie-Cités, ISCPiF*.
- BANOS, Arnaud et Cyrille GENRE-GRANDPIERRE (2012). "Towards new metrics for urban road networks : Some preliminary evidence from agent-based simulations". In : *Agent-based models of geographical systems*. Springer, p. 627–641.
- BAPTISTE, Hervé (1999). "Interactions entre le système de transport et les systèmes de villes : perspective historique pour une modélisation dynamique spatialisée". Thèse de doct. Centre d'études supérieures de l'aménagement (Tours).
- (2010). "Modeling the Evolution of a Transport System and its Impacts on a French Urban System". In : *Graphs and Networks : Multilevel Modeling, Second Edition*, p. 67–89.
- BARABASI, Albert-Laszlo (2002). "Linked : How everything is connected to everything else and what it means". In : *Plume Editors*.
- BARNDORFF-NIELSEN, Ole E et al. (2011). "Multivariate realised kernels : consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading". In : *Journal of Econometrics* 162, p. 149–169.
- BARRICO, C. et C.H. ANTUNES (2006). "Robustness Analysis in Multi-Objective Optimization Using a Degree of Robustness Concept". In : *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, p. 1887–1892. DOI : [10.1109/CEC.2006.1688537](https://doi.org/10.1109/CEC.2006.1688537).
- BARTHÉLEMY, Marc et Alessandro FLAMMINI (2008). "Modeling urban street patterns". In : *Physical review letters* 100.13, p. 138702.
- (2009). "Co-evolution of density and topology in a simple model of city formation". In : *Networks and spatial economics* 9.3, p. 401–425.
- BARTHELEMY, Marc et al. (2013). "Self-organization versus top-down planning in the evolution of a city". In : *Scientific reports* 3.
- BATTY, Michael (2007). *Cities and complexity : understanding cities with cellular automata, agent-based models, and fractals*. The MIT press.
- (2013a). "Big data, smart cities and city planning". In : *Dialogues in Human Geography* 3.3, p. 274–279.
- (2013b). *The new science of cities*. Mit Press.
- (2016). "Theoretical filters : Reducing explanations in cities to their very essence". In : *Environment and Planning B : Planning and Design* 43.5, p. 797–799.
- BATTY, Michael et Paul A LONGLEY (1994). *Fractal cities : a geometry of form and function*. Academic Press.
- BAVOUX, Jean-Jacques et al. (2005). *Géographie des transports*. Paris.
- BAZIN, Sylvie et al. (2011). "Grande vitesse ferroviaire et développement économique local : une revue de la littérature". In : *Recherche Transports Sécurité* 27.3, p. 215–238.

- BEAUCIRE, Francis et Mathieu DREVELLE (2013). "«Grand Paris Express» : un projet au service de la réduction des inégalités d'accèsibilité entre l'Ouest et l'Est de la région urbaine de Paris?" In : *Revue d'Économie Régionale & Urbaine* 3, p. 437–460.
- BEDAU, Mark (2002). "Downward causation and the autonomy of weak emergence". In : *Principia : an international journal of epistemology* 6.1, p. 5–50.
- BELMONTE, Mylène et al. (2008). "Automatisation intégrale de la ligne 1 : étude et modélisation du trafic mixte". In : *Lambda-Mu*, Session-5B.
- BENGUIGUI, Lucien et Efrat BLUMENFELD-LIEBERTHAL (2007). "A dynamic model for city size distribution beyond Zipf's law". In : *Physica A : Statistical Mechanics and its Applications* 384.2, p. 613–627.
- BENNETT, Jonathan (2010). *OpenStreetMap*. Packt Publishing Ltd.
- BERGEAUD, Antonin et al. (2017). "Classifying patents based on their semantic content". In : *PLOS ONE* 12.4, p. 1–22. DOI : [10.1371/journal.pone.0176310](https://doi.org/10.1371/journal.pone.0176310). URL : <https://doi.org/10.1371/journal.pone.0176310>.
- BERNE, Laurence (2008). "Ouverture et fermeture de territoire par les réseaux de transports dans trois espaces montagnards (Bugey, Bauges et Maurienne)". Thèse de doct. Université de Savoie.
- BERRY, Brian JL (1964). "Cities as systems within systems of cities". In : *Papers in Regional Science* 13.1, p. 147–163.
- BETTENCOURT, L. M. A. et J. LOBO (2015). "Urban Scaling in Europe". In : *ArXiv e-prints*. arXiv : [1510.00902 \[physics.soc-ph\]](https://arxiv.org/abs/1510.00902).
- BETTENCOURT, Luís MA et al. (2008). "Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics". In : *The European Physical Journal B-Condensed Matter and Complex Systems* 63.3, p. 285–293.
- BETTENCOURT, Luís MA et al. (2007). "Growth, innovation, scaling, and the pace of life in cities". In : *Proceedings of the national academy of sciences* 104.17, p. 7301–7306.
- BIRD, Steven (2006). "NLTK : the natural language toolkit". In : *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, p. 69–72.
- BLEI, David M et al. (2003). "Latent dirichlet allocation". In : *Journal of machine Learning research* 3.Jan, p. 993–1022.
- BLOOM, Nicholas et al. (2013). "Identifying Technology Spillovers and Product Market Rivalry". In : *Econometrica* 81.4, p. 1347–1393. URL : <https://ideas.repec.org/a/ecm/emetrp/v81y2013i4p1347-1393.html>.
- BLUMENFELD-LIEBERTHAL, Efrat et Juval PORTUGALI (2010). "Network cities : A complexity-network approach to urban dynamics and development". In : *Geospatial Analysis and Modelling of Urban Structure and Dynamics*. Springer, p. 77–90.

- BOLLEN, Johan et al. (2014). "From funding agencies to scientific agency". In : *EMBO reports* 15.2, p. 131–133.
- BOLÓN-CANEDO, Verónica et al. (2013). "A review of feature selection methods on synthetic data". In : *Knowledge and information systems* 34.3, p. 483–519.
- BOLOTIN, A. (2014). "Computational solution to quantum foundational problems". In : *ArXiv e-prints*. arXiv : 1403.7686 [quant-ph].
- BONANNO, G. et al. (2001). "Levels of complexity in financial markets". In : *Physica A Statistical Mechanics and its Applications* 299, p. 16–27. eprint : cond-mat/0104369.
- BONIN, Olivier, Jean-Paul HUBERT et al. (2012). "Modèle de morphogénèse urbaine : simulation d'espaces qualitativement différenciés dans le cadre du modèle de l'économie urbaine". In : *49è colloque de l'ASRDLF*.
- BONNAFOUS, Alain et François PLASSARD (1974). "Les méthodologies usuelles de l'étude des effets structurants de l'offre de transport". In : *Revue économique*, p. 208–232.
- BONNAFOUS, Alain et al. (1974). "La detection des effets structurants d'autoroute : Application à la Vallée du Rhône". English. In : *Revue économique* 25.2, pp. 233–256. ISSN : 00352764. URL : <http://www.jstor.org/stable/3500568>.
- BOSCH, F van den et al. (1990). "The velocity of spatial population expansion". In : *Journal of Mathematical Biology* 28.5, p. 529–565.
- BOUCHAUD, J. P. et M. POTTERS (2009). "Financial Applications of Random Matrix Theory : a short review". In : *ArXiv e-prints*. arXiv : 0910.1205 [q-fin.ST].
- BOUCHAUD, J-P et al. (2000). "Apparent multifractality in financial time series". In : *The European Physical Journal B-Condensed Matter and Complex Systems* 13.3, p. 595–599.
- BOURGINE, P. et al. (2009). "French Roadmap for complex Systems 2008-2009". In : *ArXiv e-prints*. arXiv : 0907.2221 [nlin.AO].
- BOURGINE, Paul et John STEWART (2004). "Autopoiesis and cognition". In : *Artificial life* 10.3, p. 327–345.
- BOUTEILLER, Catherine et Sybille BERJOAN (2013). "Open data en transport urbain : quelles sont les données mises à disposition ? Quelles sont les stratégies des autorités organisatrices ?" In :
- BRAND, Christian et al. (2013). "Accelerating the transformation to a low carbon passenger transport system : The role of car purchase taxes, feebates, road taxes and scrappage incentives in the UK". In : *Transportation Research Part A : Policy and Practice* 49, p. 132–148.
- BRETAGNOLLE, Anne (2009a). "Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis". Français. HDR. Université Panthéon-Sorbonne - Paris I. URL : <http://tel.archives-ouvertes.fr/tel-00459720>.

- BRETAGNOLLE, Anne (2009b). "Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis". Français. HDR. Université Panthéon-Sorbonne - Paris I. URL : <http://tel.archives-ouvertes.fr/tel-00459720>.
- BRETAGNOLLE, Anne et al. (2006). "From theory to modelling : urban systems as complex systems". In : *CyberGeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et al. (2002). "Time and space scales for measuring urban growth". In : *Cybergeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et Denise PUMAIN (2010a). "Comparer deux types de systèmes de villes par la modélisation multi-agents". In : *Qu'appelle t-on aujourd'hui les sciences de la complexité ? Langages, réseaux, marchés, territoires*, p. 271–299.
- (2010b). "Simulating Urban Networks through Multiscalar Space-Time Dynamics : Europe and the United States, 17th-20th Centuries". In : *Urban Studies* 47.13, p. 2819–2839. DOI : [10.1177/0042098010377366](https://doi.org/10.1177/0042098010377366). eprint : <http://dx.doi.org/10.1177/0042098010377366>. URL : <http://dx.doi.org/10.1177/0042098010377366>.
- BRETAGNOLLE, Anne et al. (1998). "Space-time contraction and the dynamics of urban systems". In : *Cybergeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et al. (2009). "The organization of urban systems". In : *Complexity perspectives in innovation and social change*. Springer, p. 197–220.
- BROWN, Matthew J (2009). "Models and perspectives on stage : remarks on Giere's scientific perspectivism". In : *Studies in History and Philosophy of Science Part A* 40.2, p. 213–220.
- BRUCK, Péter et al. (2016). "Recognition of emerging technology trends : class-selective study of citations in the US Patent Citation Network". In : *Scientometrics* 107.3, p. 1465–1475.
- BRUNSDON, Chris et al. (1996). "Geographically weighted regression : a method for exploring spatial nonstationarity". In : *Geographical analysis* 28.4, p. 281–298.
- BRUNSDON, Chris et al. (1998). "Geographically weighted regression". In : *Journal of the Royal Statistical Society : Series D (The Statistician)* 47.3, p. 431–443.
- BULCKE, Tim Van den et al. (2006). "SynTReN : a generator of synthetic gene expression data for design and analysis of structure learning algorithms". In : *BMC bioinformatics* 7.1, p. 43.
- CAMERER, Colin F et al. (2016). "Evaluating replicability of laboratory experiments in economics". In : *Science*, aaf0918.
- CARLILE, Paul R (2004). "Transferring, translating, and transforming : An integrative framework for managing knowledge across boundaries". In : *Organization science* 15.5, p. 555–568.

- CARVER, Stephen J (1991). "Integrating multi-criteria evaluation with geographical information systems". In : *International Journal of Geographical Information System* 5.3, p. 321–339.
- CERQUEIRA, Eugênia Viana (2017). "Les inégalités d'accès aux ressources urbaines dans les franges périphériques de Belo Horizonte (Brésil) : quelles évolutions ?" In : *EchoGéo* 39.
- CHALIDABHONGSE, Junavit et CC Jay Kuo (1997). "Fast motion vector estimation using multiresolution-spatio-temporal correlations". In : *Circuits and Systems for Video Technology, IEEE Transactions on* 7.3, p. 477–488.
- CHAMPOILLION, Pierre (2006). "TERRITORY AND TERRITORIALIZATION : PRESENT STATE OF THE CAENTI THOUGHT". In : *International Conference of Territorial Intelligence*. INTI-International Network of Territorial Intelligence. Alba Iulia, Romania, p51–58. URL : <https://halshs.archives-ouvertes.fr/halshs-00999026>.
- CHANG, Justin S (2006). "Models of the Relationship between Transport and Land-use : A Review". In : *Transport Reviews* 26.3, p. 325–350.
- CHASSET, Pierre-Olivier et al. (2016). *cybergeo20 v1.0*. DOI : [10.5281/zenodo.53905](https://doi.org/10.5281/zenodo.53905). URL : <http://dx.doi.org/10.5281/zenodo.53905>.
- CHAVALARIAS, David (2016). "What's wrong with Science ?" In : *Scientometrics*, p. 1–23.
- CHAVALARIAS, David et Jean-Philippe COINTET (2013). "Phylomemetic patterns in science evolution—the rise and fall of scientific fields". In : *Plos One* 8.2, e54847.
- CHAVALARIAS, David et al. (2005). "Nobel, Le Jeu De La Découverte Scientifique". In :
- CHEN, Duan-Rung et Khoa TRUONG (2012). "Using multilevel modeling and geographically weighted regression to identify spatial variations in the relationship between place-level disadvantages and obesity in Taiwan". In : *Applied Geography* 32.2, p. 737–745.
- CHEN, Wenling et David M LEVINSON (2006). "Effectiveness of learning transportation network growth through simulation". In : *Journal of Professional Issues in Engineering Education and Practice* 132.1, p. 29–41.
- CHEN, Yanguang (2009). "Urban gravity model based on cross-correlation function and Fourier analyses of spatio-temporal process". In : *Chaos, Solitons & Fractals* 41.2, p. 603–614.
- (2010). "Characterizing growth and form of fractal cities with allometric scaling exponents". In : *Discrete Dynamics in Nature and Society* 2010.
- CHÉREL, Guillaume et al. (2015). "Beyond Corroboration : Strengthening Model Validation by Looking for Unexpected Patterns". In : *PLoS ONE* 10.9, e0138212. DOI : [10.1371/journal.pone.0138212](https://doi.org/10.1371/journal.pone.0138212). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0138212>.

- CHICHEPORTICHE, Rémy et Jean-Philippe BOUCHAUD (2013). "A nested factor model for non-linear dependences in stock returns". In : *arXiv preprint arXiv :1309.3102*.
- CHOI, Jinho et Yong-Sik HWANG (2014). "Patent keyword network analysis for improving technology development efficiency". In : *Technological Forecasting and Social Change* 83, p. 170–182.
- CLAUSSET, Aaron et al. (2004). "Finding community structure in very large networks". In : *Physical review E* 70.6, p. 066111.
- CLAVAL, Paul (1985). "Causalité et géographie". In : *Espace géographique* 14.2, p. 109–115.
- (1987). "Réseaux territoriaux et enracinement". In : *Cahier/Groupe Réseaux* 3.7, p. 44–60.
- COLANDER, David (2003). *The complexity revolution and the future of economics*. Rapp. tech. Middlebury College, Department of Economics.
- COMBES, Pierre-Philippe et Miren LAFOURCADE (2005). "Transport costs : measures, determinants, and regional policy implications for France". In : *Journal of Economic Geography* 5.3, p. 319–349.
- COMMENGES, H (2013a). "The invention of daily mobility : Performative aspects of the instruments of economics of transportation." In : *Theses, Université Paris-Diderot-Paris VII*.
- COMMENGES, Hadrien (2013b). "The invention of daily mobility. Performative aspects of the instruments of economics of transportation." Theses. Université Paris-Diderot - Paris VII. URL : <https://tel.archives-ouvertes.fr/tel-00923682>.
- COTTINEAU, C. (2016). "MetaZipf. (Re)producing knowledge about city size distributions". In : *ArXiv e-prints*. arXiv : [1606 . 06162 \[physics.soc-ph\]](https://arxiv.org/abs/1606.06162).
- COTTINEAU, C et al. (2016). "Back to the Future of Multimodeling". In : *Royal Geographical Society-Annual Conference 2016-Session : Geocomputation, the Next 20 Years* (1).
- COTTINEAU, Clementine (2014). "L'évolution des villes dans l'espace post-soviétique. Observation et modélisations." Thèse de doct. Université Paris 1 Panthéon-Sorbonne.
- COTTINEAU, Clémentine (2015). *Urban scaling : What cities are we talking about?* Presentation of ongoing work at Quanturb seminar, April 1st 2015.
- COTTINEAU, Clémentine et al. (2015). "An incremental method for building and evaluating agent-based models of systems of cities". In :
- COTTINEAU, Clémentine et al. (2015a). "A modular modelling framework for hypotheses testing in the simulation of urbanisation". In : *Systems* 3.4, p. 348–377.
- COTTINEAU, Clémentine et al. (2015b). "Revisiting some geography classics with spatial simulation". In : *Plurimondi. An International Forum for Research and Debate on Human Settlements*. T. 7. 15.

- COTTINEAU, Clementine et al. (2017). "Initial spatial conditions in simulation models : the missing leg of sensitivity analyses?" In : *Geocomputation Conference*.
- COURTAT, Thomas et al. (2011). "Mathematics and morphogenesis of cities : A geometrical approach". In : *Physical Review E* 83.3, p. 036106.
- CRONIN, Blaise et Cassidy R SUGIMOTO (2014). *Beyond bibliometrics : Harnessing multidimensional indicators of scholarly impact*. MIT Press.
- CROSATO, Emanuele et al. (2017). "Informative and misinformative interactions in a school of fish". In : *arXiv preprint arXiv :1705.01213*.
- CROSS, MC, PC HOHENBERG et al. (1994). "Spatiotemporal chaos". In : *Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information* 263.5153, p. 1569–1569.
- CROZET, Yves et Francois DUMONT (2011a). "Retour sur les effets économiques du TGV. Les effets structurants sont un mythe (interview)". In : *Ville, Rail et Transports* 525, p. 48–51. URL : <https://halshs.archives-ouvertes.fr/halshs-01094554>.
- (2011b). "Retour sur les effets économiques du TGV. Les effets structurants sont un mythe (interview)". In : *Ville, Rail et Transports* 525, p. 48–51. URL : <https://halshs.archives-ouvertes.fr/halshs-01094554>.
- CURRAN, Clive-Steven et Jens LEKER (2011). "Patent indicators for monitoring convergence—examples from NFF and ICT". In : *Technological Forecasting and Social Change* 78.2, p. 256–273.
- CUYALA, Sylvain (2014). "Analyse spatio-temporelle d'un mouvement scientifique. L'exemple de la géographie théorique et quantitative européenne francophone." Thèse de doct. Université Paris 1 Panthéon-Sorbonne.
- DAMM, David et al. (1980). "Response of urban real estate values in anticipation of the Washington Metro". In : *Journal of Transport Economics and Policy*, p. 315–336.
- DE DOMENICO, Manlio et al. (2015). "Ranking in interconnected multilayer networks reveals versatile nodes". In : *Nature communications* 6.
- DE LEON, FD et al. (2007). "NetLogo Urban Suite-Tijuana Border-towns model". In : *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*. URL : <http://ccl.northwestern.edu/netlogo/models/UrbanSuite-TijuanaBordertowns>.
- DE NADAI, Marco et al. (2016). "The Death and Life of Great Italian Cities : A Mobile Phone Data Perspective". In : *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, p. 413–423.
- DEB, Kalyanmoy et Himanshu GUPTA (2006). "Introducing robustness in multi-objective optimization". In : *Evolutionary Computation* 14.4, p. 463–494.

- DECELLE, Aurelien et al. (2011). "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In : *Physical Review E* 84.6, p. 066106.
- DECHEZLEPRÂTRE, Antoine et al. (2014). *Knowledge Spillovers from Clean and Dirty Technologies*. CEP Discussion Papers dp1300. Centre for Economic Performance, LSE. url : <https://ideas.repec.org/p/cep/cepdps/dp1300.html>.
- DEFFUANT, Guillaume et al. (2015). "Visions de la complexité. Le démon de Laplace dans tous ses états". In : *Natures Sciences Sociétés* 23.1, p. 42–53.
- DELILE, Julien et al. (2016). "Chapitre 17. Modélisation multi-agent de l'embryogenèse animale". In : *Modélisations, simulations, systèmes complexes*, p. 581–624.
- DELONS, Jean et al. (2008). "PIRANDELLO an integrated transport and land-use model for the Paris area". URL : <https://halv3-preprod.archives-ouvertes.fr/hal-00319087>.
- DESJARDINS, Xavier (2010). "la bataille du Grand Paris". In : *L'Information géographique* 74.4, p. 29–46.
- DI MEO, Guy (1998). "De l'espace aux territoires : éléments pour une archéologie des concepts fondamentaux de la géographie". In : *L'information géographique* 62.3, p. 99–110.
- DICK, Josef et Friedrich PILLICHSHAMMER (2010). *Digital nets and sequences : Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press.
- DIDEROT, Denis (1965). *Entretien entre d'Alembert et Diderot*. Garnier-Flammarion.
- DIRK, Lynn (1999). "A Measure of Originality The Elements of Science". In : *Social Studies of Science* 29.5, p. 765–776.
- DOBBIE, Melissa J et David DAIL (2013). "Robustness and sensitivity of weighting and aggregation in constructing composite indices". In : *Ecological Indicators* 29, p. 270–277.
- DOLLFUS, O et F Durand DASTÈS (1975). "Some remarks on the notions of 'structure' and 'system' in geography". In : *Geoforum* 6.2, p. 83–94.
- DOURSAT, René et al. (2012). *Morphogenetic engineering : toward programmable complex systems*. Springer.
- DRAGOMIR, SS (1999). "The Ostrowski's integral inequality for Lipschitzian mappings and applications". In : *Computers & Mathematics with Applications* 38.11, p. 33–37.
- DRUMMOND, Chris (2009). "Replicability is not reproducibility : nor is it good science". In :
- DUCRUET, César et Laurent BEAUGUITTE (2014). "Spatial science and network science : Review and outcomes of a complex relationship". In : *Networks and Spatial Economics* 14.3-4, p. 297–316.
- DUDA, John (2013). "Cybernetics, anarchism and self-organisation". In : *Anarchist studies* 21.1, p. 52.

- DUPUY, Gabriel (1987). "Vers une théorie territoriale des réseaux : une application au transport urbain". In : *Annales de Géographie*. JS-TOR, p. 658–679.
- DUPUY, Gabriel et Lucien Gilles BENGUIGUI (2015). "Sciences urbaines : interdisciplinarités passive, naïve, transitive, offensive". In : *Métropoles* 16.
- DURAND-DASTES, François (2003). *Les géographes et la notion de causalité*.
- DURANTIN, Arnaud et al. (2017). "Disruptive Innovation in Complex Systems". In : *Complex Systems Design & Management*. Springer, p. 41–56.
- DURANTON, Gilles (1999). "Distance, land, and proximity : economic analysis and the evolution of cities". In : *Environment and Planning a* 31.12, p. 2169–2188.
- EPSTEIN, Joshua M et Robert L AXTELL (1996). *Growing artificial societies : Social science from the bottom up (complex adaptive systems)*. Brookings Institution Press MIT Press.
- FARMER, J Doyne et Duncan FOLEY (2009). "The economy needs agent-based modelling". In : *Nature* 460.7256, p. 685–686.
- FATTORI, Michele et al. (2003). "Text mining applied to patent mapping : a practical business case". In : *World Patent Information* 25.4, p. 335–342.
- FAVARO, Jean-Marc et Denise PUMAIN (2011). "Gibrat Revisited : An Urban Growth Model Incorporating Spatial Interaction and Innovation Cycles." In : *Geographical Analysis* 43.3, p. 261–286.
- FEBRES, Gerardo et al. (2013). "Complexity measurement of natural and artificial languages". In : *arXiv preprint arXiv :1311.5427*.
- FEYERABEND, Paul (1993). *Against method*. Verso.
- FOOT, Robin (1994). "RATP, un corporatisme à l'épreuve des voyageurs". In : *Travail* 31, p. 63–100.
- (2005). "Faut-il protéger le métro des voyageurs ? Ou l'appréhension du voyageur par les ingénieurs et les conducteurs". In : *Tra vailler* 2, p. 169–206.
- FRANCO, Jessica et al. (2009). "DiceDesign-package". In : *Designs of Computer Experiments*, p. 2.
- FRANK, Morgan R et al. (2014). "Constructing a taxonomy of fine-grained human movement and activity motifs through social media". In : *arXiv preprint arXiv :1410.1393*.
- FRANKHAUSER, Pierre (1998). "Fractal geometry of urban patterns and their morphogenesis". In : *Discrete Dynamics in Nature and Society* 2.2, p. 127–145.
- (2008). "Fractal geometry for measuring and modelling urban patterns". In : *The dynamics of complex urban systems*. Springer, p. 213–243.

- FRIGG, Roman et Ioannis VOTSIS (2011). "Everything you always wanted to know about structural realism but were afraid to ask". In : *European journal for philosophy of science* 1.2, p. 227–276.
- FUJITA, Masahisa et al. (1999). "On the evolution of hierarchical urban systems". In : *European Economic Review* 43.2, p. 209–251.
- FULLERTON, Don et Sarah E WEST (2002). "Can taxes on cars and on gasoline mimic an unavailable tax on emissions?" In : *Journal of Environmental Economics and Management* 43.1, p. 135–157.
- FURMAN, Jeffrey L. et Scott STERN (2011). "Climbing atop the Shoulders of Giants : The Impact of Institutions on Cumulative Research". In : *American Economic Review* 101.5, p. 1933–63. DOI : [10.1257/aer.101.5.1933](https://doi.org/10.1257/aer.101.5.1933). URL : <http://www.aeaweb.org/articles?id=10.1257/aer.101.5.1933>.
- FUSCO, Giovanni (2004). "La mobilité quotidienne dans les grandes villes du monde : application de la théorie des réseaux bayésiens". In : *Cybergeo : European Journal of Geography*.
- GABAIX, Xavier (1999). "Zipf's law for cities : an explanation". In : *Quarterly journal of Economics*, p. 739–767.
- GABAIX, Xavier et Yannis M. IOANNIDES (2004). "Chapter 53 The evolution of city size distributions". In : *Cities and Geography*. Sous la dir. de J. Vernon HENDERSON et Jacques-François THISSE. T. 4. Handbook of Regional and Urban Economics. Elsevier, p. 2341–2378. DOI : [http://dx.doi.org/10.1016/S1574-0080\(04\)80010-5](https://doi.org/10.1016/S1574-0080(04)80010-5). URL : <http://www.sciencedirect.com/science/article/pii/S1574008004800105>.
- GABORA, L. et M. STEEL (2017). "Autocatalytic networks in cognition and the origin of culture". In : *ArXiv e-prints*. arXiv : [1703.05917 \[q-bio.NC\]](https://arxiv.org/abs/1703.05917).
- GAO, Zhong-Ke et al. (2017). "Complex network analysis of time series". In : *EPL (Europhysics Letters)* 116.5, p. 50001.
- GAO, Zhong-Ke et al. (2015). "Multiscale complex network for analyzing experimental multivariate time series". In : *EPL (Europhysics Letters)* 109.3, p. 30005.
- GAUTIER, Erwan et Ronan Le SAOUT (2015). "The dynamics of gasoline prices : Evidence from daily French micro data". In : *Journal of Money, Credit and Banking* 47.6, p. 1063–1089.
- GELL-MANN, Murray (1995). *The Quark and the Jaguar : Adventures in the Simple and the Complex*. Macmillan.
- GEMINO, Andrew et Yair WAND (2004). "A framework for empirical evaluation of conceptual modeling techniques". In : *Requirements Engineering* 9.4, p. 248–260.
- GERKEN, Jan M et Martin G MOEHRLE (2012). "A new instrument for technology monitoring : novelty in patents measured by semantic patent analysis". In : *Scientometrics* 91.3, p. 645–670.
- GIERE, Ronald N (2010a). "An agent-based conception of models and scientific representation". In : *Synthese* 172.2, p. 269–281.

- (2010b). *Explaining science : A cognitive approach*. University of Chicago Press.
- (2010c). *Scientific perspectivism*. University of Chicago Press.
- GILLI, Frédéric et Jean-Marc OFFNER (2009). *Paris, métropole hors les murs : aménager et gouverner un Grand Paris*. Sciences Po, les presses.
- GIRRES, Jean-François et Guillaume TOUYA (2010). “Quality assessment of the French OpenStreetMap dataset”. In : *Transactions in GIS* 14.4, p. 435–459.
- GLEYZE, Jean-François (2005). “La vulnérabilité structurelle des réseaux de transport dans un contexte de risques”. Thèse de doct. Université Paris-Diderot-Paris VII.
- GOLDEN, Boris et al. (2012). “Modeling of complex systems ii : A minimalist and unified semantics for heterogeneous integrated systems”. In : *Applied Mathematics and Computation* 218.16, p. 8039–8055.
- GOLDHIRSCH, Isaac et al. (1987). “Stability and Lyapunov stability of dynamical systems : A differential approach and a numerical method”. In : *Physica D : Nonlinear Phenomena* 27.3, p. 311–337.
- GOTTMANN, Jean (1964). *Megalopolis : the urbanized northeastern seaboard of the United States*. MIT Press Cambridge, MA.
- GREGG, Jay S et al. (2009). “The temporal and spatial distribution of carbon dioxide emissions from fossil-fuel use in North America”. In : *Journal of Applied Meteorology and Climatology* 48.12, p. 2528–2542.
- GRIFFITH, Daniel A (1980). “Towards a theory of spatial statistics”. In : *Geographical Analysis* 12.4, p. 325–339.
- (1992). “What is spatial autocorrelation ? Reflections on the past 25 years of spatial statistics”. In : *Espace géographique* 21.3, p. 265–280.
- (2012). *Advanced spatial statistics : special topics in the exploration of quantitative spatial data series*. T. 12. Springer Science & Business Media.
- GRILICHES, Zvi (1990). *Patent Statistics as Economic Indicators : A Survey*. NBER Working Papers 3301. National Bureau of Economic Research, Inc. URL : <https://ideas.repec.org/p/nbr/nberwo/3301.html>.
- GRIMM, Volker et al. (2005). “Pattern-oriented modeling of agent-based complex systems : lessons from ecology”. In : *science* 310.5750, p. 987–991.
- GUENÉE, Bernard (1968). “Espace et Etat dans la France du bas Moyen Age”. In : *Annales. Histoire, Sciences Sociales*. T. 23. 4. JSTOR, p. 744–758.
- GUÉROIS, Marianne et Renaud LE GOIX (2009). “La dynamique spatio-temporelle des prix immobiliers à différentes échelles : le cas des appartements anciens à Paris (1990-2003)”. In : *Cybergeo : European Journal of Geography*.

- GUÉROIS, Marianne et Fabien PAULUS (2002). "Commune centre, agglomération, aire urbaine : quelle pertinence pour l'étude des villes?" In : *Cybergeo : European Journal of Geography*.
- GUO, Xiaolei et Henry X LIU (2011). "Bounded rationality and irreversible network change". In : *Transportation Research Part B : Methodological* 45.10, p. 1606–1618.
- GURCIULLO, S. et al. (2015). "Complex Politics : A Quantitative Semantic and Topological Analysis of UK House of Commons Debates". In : *ArXiv e-prints*. arXiv : 1510.03797 [physics.soc-ph].
- GUTMANN, Amy (2011). "The ethics of synthetic biology : guiding principles for emerging technologies". In : *Hastings Center Report* 41.4, p. 17–22.
- HACHI, R. (2013). *La fractalité comme indicateur de l'état de conservation du patrimoine urbain, Master Thesis Memoire*. Rapp. tech. Université Paris VII.
- HACKING, Ian (1999). *The social construction of what?* Harvard university press.
- HAKEN, Herman et Juval PORTUGALI (2003). "The face of the city is its information". In : *Journal of Environmental Psychology* 23.4, p. 385–408.
- HAKEN, Hermann (1980). "Synergetics". In : *Naturwissenschaften* 67.3, p. 121–128.
- HAKLAY, Mordechai (2010). "How good is volunteered geographical information ? A comparative study of OpenStreetMap and Ordnance Survey datasets". In : *Environment and planning B : Planning and design* 37.4, p. 682–703.
- HALL, Bronwyn H et al. (2001). *The NBER Patent Citations Data File : Lessons, Insights and Methodological Tools*. CEPR Discussion Papers 3094. C.E.P.R. Discussion Papers. URL : <https://ideas.repec.org/p/cpr/ceprdp/3094.html>.
- HALL, Peter Geoffrey et Kathy PAIN (2006). *The polycentric metropolis : learning from mega-city regions in Europe*. Routledge.
- HAMERLY, Greg, Charles ELKAN et al. (2003). "Learning the k in k-means". In : *NIPS*. T. 3, p. 281–288.
- HAN, Sangjin (2003). "Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks". In : *Transportation Research Part B : Methodological* 37.3, p. 225–249.
- HARAN, EGP et Daniel R VINING (1973). "A MODIFIED YULE-SIMON MODEL ALLOWING FOR INTERCITY MIGRATION AND ACCOUNTING FOR THE OBSERVED FORM OF THE SIZE DISTRIBUTION OF CITIES*". In : *Journal of Regional Science* 13.3, p. 421–437.
- HATCHUEL, Armand et al. (1988). "Des stations de métro en mouvement : Station 2000, un scénario prospectif". In : *Les Annales de la recherche urbaine*. T. 39. 1. Persée-Portail des revues scientifiques en SHS, p. 35–42.

- HATTON, L. et G. WARR (2016). "Full Computational Reproducibility in Biological Science : Methods, Software and a Case Study in Protein Biology". In : *ArXiv e-prints*. arXiv : [1608.06897 \[q-bio.QM\]](https://arxiv.org/abs/1608.06897).
- HIDALGO, C. A. (2015). "Disconnected! The parallel streams of network literature in the natural and social sciences". In : *ArXiv e-prints*. arXiv : [1511.03981 \[physics.soc-ph\]](https://arxiv.org/abs/1511.03981).
- HILLIER, Bill (2016). "The Fourth Sustainability, Creativity : Statistical Associations and Credible Mechanisms". In : *Complexity, Cognition, Urban Planning and Design*. Springer, p. 75–92.
- HILLIER, Bill et Julienne HANSON (1989). *The social logic of space*. Cambridge university press.
- HOFSTADTER, Douglas H (1980). *Gödel, Escher, Bach : An Eternal Golden Braid ;[a Metaphoric Fugue on Minds and Machines in the Spirit of Lewis Carroll]*. Penguin Books.
- HOLLAND, John H (2012). *Signals and boundaries : Building blocks for complex adaptive systems*. Mit Press.
- HOLMES, C. et al. (2017). "Luria-Delbrück, revisited : The classic experiment does not rule out Lamarckian evolution". In : *ArXiv e-prints*. arXiv : [1701.05627 \[q-bio.PE\]](https://arxiv.org/abs/1701.05627).
- HOLZINGER, Andreas et al. (2014). "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions". In : *BMC bioinformatics* 15.6, p. I1.
- HOMOCIANU, Marius (2009). "Transport-land use interaction modeling - Residential choices of households in urban area of Lyon". Theses. Université Lumière - Lyon II. URL : <https://tel.archives-ouvertes.fr/tel-00359302>.
- HUNG, Ling-Hong et al. (2016). "GUIDock : Using Docker Containers with a Common Graphics User Interface to Address the Reproducibility of Research". In : *PLoS ONE* 11.4, p. 1–14. DOI : [10.1371/journal.pone.0152686](https://doi.org/10.1371/journal.pone.0152686). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0152686>.
- Hypergeo*. <http://www.hypergeo.eu/spip.php?page=sommaire>.
- IACONO, Michael et al. (2008). "Models of transportation and land use change : a guide to the territory". In : *Journal of Planning Literature* 22.4, p. 323–340.
- IACOVACCI, Jacopo et al. (2015). "Mesoscopic Structures Reveal the Network Between the Layers of Multiplex Datasets". In : *arXiv preprint arXiv :1505.03824*.
- JARROW, Robert A (1999). "In Honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes : A Partial Differential Equation that Changed the World". In : *The Journal of Economic Perspectives*, p. 229–248.
- JÉGOU, Anne et al. (2012). "L'évaluation par indicateurs : un outil nécessaire d'aménagement urbain durable ?. Réflexions à partir de la démarche parisienne pour le géographe et l'aménageur". In : *Cybergeo : European Journal of Geography*.

- KAPLAN, Sarah et Keyvan VAKILI (2015). "The double-edged sword of recombination in breakthrough innovation". In : *Strategic Management Journal* 36.10, p. 1435–1457.
- KARATZOGLOU, Alexandros et al. (2004). "kernlab – An S4 Package for Kernel Methods in R". In : *Journal of Statistical Software* 11.9, p. 1–20. URL : <http://www.jstatsoft.org/v11/i09/>.
- KATZ, Michael L (1996). "Remarks on the economic implications of convergence". In : *Industrial and Corporate Change* 5.4, p. 1079–1095.
- KAY, Luciano et al. (2014). "Patent overlay mapping : Visualizing technological distance". In : *Journal of the Association for Information Science and Technology* 65.12, p. 2432–2443.
- KE, Yan et al. (2007). "Spatio-temporal shape and flow correlation for action recognition". In : *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, p. 1–8.
- KEERSMAECKER, Marie-Laurence et al. (2003). "Using fractal dimensions for characterizing intra-urban diversity : The example of Brussels". In : *Geographical analysis* 35.4, p. 310–328.
- KNIGHT, Frank B (1975). "A predictive view of continuous time processes". In : *The annals of Probability*, p. 573–596.
- KOCH, Christof et Gilles LAURENT (1999). "Complexity and the nervous system". In : *Science* 284.5411, p. 96–98.
- KOCH, Julian et Simon STISEN (2017). "Citizen science : A new perspective to advance spatial pattern evaluation in hydrology". In : *PLOS ONE* 12.5, p. 1–20. DOI : [10.1371/journal.pone.0178165](https://doi.org/10.1371/journal.pone.0178165). URL : <https://doi.org/10.1371/journal.pone.0178165>.
- KOLCHINSKY, A. et al. (2015). "Modularity and the Spread of Perturbations in Complex Dynamical Systems". In : *ArXiv e-prints*. arXiv : [1509.04386 \[physics.soc-ph\]](https://arxiv.org/abs/1509.04386).
- KRYVOBOKOV, Marko et al. (2013). "Comparison of Static and Dynamic Land Use-Transport Interaction Models". In : *Transportation Research Record : Journal of the Transportation Research Board* 2344.1, p. 49–58.
- KUHN, Thomas S (1970). *The structure of scientific revolutions*.
- KWAN, M.P. (2012). "The uncertain geographic context problem". In : *Annals of the Association of American Geographers* 102.5, p. 958–968.
- KWAN, Mei-Po (1998). "Space-time and integral measures of individual accessibility : a comparative analysis using a point-based framework". In : *Geographical analysis* 30.3, p. 191–216.
- L'ESPACE GÉOGRAPHIQUE (2014). *Les effets structurants des infrastructures de transport, L'Espace géographique 2014/1 (Tome 43)*, p. 51–67.
- L'HOSTIS, Alain et al. (2012). "La ville orientée vers le rail". In : *Ville et mobilité*.
- LAGESSE, C. (2015). "Read Cities through their Lines. Methodology to characterize spatial graphs". In : *ArXiv e-prints*. arXiv : [1512.01268 \[physics.soc-ph\]](https://arxiv.org/abs/1512.01268).

- LAUGHLIN, Robert B (2006). *A different universe : Reinventing physics from the bottom down*. Basic Books.
- LAUNER, Robert L et Graham N WILKINSON (2014). *Robustness in statistics*. Academic Press.
- LE NÉCHET, Florent (2015). "De la forme urbaine à la structure métropolitaine : une typologie de la configuration interne des densités pour les principales métropoles européennes de l'audit urbain". In : *Cybergeo : European Journal of Geography*. URL : <http://cybergeo.revues.org/26753> (visité le 21/04/2015).
- LE NÉCHET, Florent (2015). "De la forme urbaine à la structure métropolitaine : une typologie de la configuration interne des densités pour les principales métropoles européennes de l'Audit Urbain". In : *Cybergeo : European Journal of Geography*.
- LE NÉCHET, Florent et Juste RAIMBAULT (2015). "Modeling the emergence of metropolitan transport authority in a polycentric urban region". In : *Plurimondi. An International Forum for Research and Debate on Human Settlements* 7.15.
- LE TEXIER, Marion et Geoffrey CARUSO (2017). "Assessing geographical effects in spatial diffusion processes : The case of euro coins". In : *Computer, Environment and Urban Systems* 61.A, p. 81–93.
- LECHNER, Thomas et al. (2004). "Procedural modeling of land use in cities". In :
- LECHNER, Thomas et al. (2006). "Procedural modeling of urban land use". In : *ACM SIGGRAPH 2006 Research posters*. ACM, p. 135.
- LEE, SeongWoo et al. (2009). "Determinants of crime incidence in Korea : a mixed GWR approach". In : *World conference of the spatial econometrics association*, p. 8–10.
- LEEUW, Sander van der et al. (2009). "The Long-Term Evolution of Social Organization". In : *Complexity Perspectives in Innovation and Social Change*. Sous la dir. de David LANE et al. Dordrecht : Springer Netherlands, p. 85–116. ISBN : 978-1-4020-9663-1. DOI : [10.1007/978-1-4020-9663-1_4](https://doi.org/10.1007/978-1-4020-9663-1_4). URL : http://dx.doi.org/10.1007/978-1-4020-9663-1_4.
- LEGAVRE, Jean Baptiste (1996). "La «neutralité» dans l'entretien de recherche. Retour personnel sur une évidence". In : *Politix* 9.35, p. 207–225.
- LERNER, Josh et Amit SERU (2015). "The use and misuse of patent data : Issues for corporate finance and beyond". In : *Booth/Harvard Business School Working Paper*.
- LEURENT, Fabien et Houda BOUJNAH (2014). "A user equilibrium, traffic assignment model of network route and parking lot choice, with search circuits and cruising flows". In : *Transportation Research Part C : Emerging Technologies* 47, p. 28–46.
- LEVINSON, David Matthew et al. (2007). "The co-evolution of land use and road networks". In : *Transportation and traffic theory*, p. 839–859.

- LEVINSON, David (2008). "Density and dispersion : the co-development of land use and rail in London". In : *Journal of Economic Geography* 8.1, p. 55–77.
- LEVINSON, David et Wei CHEN (2005). "Paving new ground : a Markov chain model of the change in transportation networks and land use". In : *Access to destinations*. Emerald Group Publishing Limited, p. 243–266.
- LEVINSON, David et Ramachandra KARAMALAPUTI (2003). "Induced supply : a model of highway network expansion at the microscopic level". In : *Journal of Transport Economics and Policy (JTEP)* 37.3, p. 297–318.
- LEVINSON, David et al. (2012). "Forecasting and evaluating network growth". In : *Networks and Spatial Economics* 12.2, p. 239–262.
- LI, Guan-Cheng et al. (2014). "Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)". In : *Research Policy* 43.6, p. 941–955.
- LI, J et U WILENSKY (2009). *NetLogo Sugarscape 3 Wealth Distribution model*.
- LISSACK, Michael (2013). "Subliminal influence or plagiarism by negligence ? The Slodderwetenschap of ignoring the internet". In : *Journal of Academic Ethics*.
- LIU, Wei et al. (2011). "Discovering spatio-temporal causal interactions in traffic data streams". In : *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 1010–1018.
- LIVET, Pierre et al. (2010). "Ontology, a Mediator for Agent-Based Modeling in Social Science". In : *Journal of Artificial Societies and Social Simulation* 13.1, p. 3. ISSN : 1460-7425. URL : <http://jasss.soc.surrey.ac.uk/13/1/3.html>.
- LOI, Daniel (1985). "Une étude de la causalité dans la géographie classique française.[L'exemple des premières thèses régionales]". In : *Espace géographique* 14.2, p. 121–125.
- LOUAIL, Thomas et al. (2016). "Crowdsourcing the Robin Hood effect in cities". In : *arXiv preprint arXiv:1604.08394*.
- LOUF, R. et M. BARTHELEMY (2014). "How congestion shapes cities : from mobility patterns to scaling". In : *ArXiv e-prints*. arXiv : 1401.8200 [physics.soc-ph].
- LOUF, Rémi et Marc BARTHELEMY (2014a). "A typology of street patterns". In : *Journal of The Royal Society Interface* 11.101, p. 20140924.
- (2014b). "Scaling : lost in the smog". In : *arXiv preprint arXiv:1410.4964*.
- (2015). "Patterns of residential segregation". In : *arXiv preprint arXiv:1511.04268*.
- LOUF, Rémi et al. (2013). "Emergence of hierarchy in cost-driven growth of spatial networks". In : *Proceedings of the National Academy of Sciences* 110.22, p. 8824–8829.

- LOUF, Rémi et al. (2014). "Scaling in Transportation Networks". In : *PLoS ONE* 9.7, e102007. DOI : [10.1371/journal.pone.0102007](https://doi.org/10.1371/journal.pone.0102007). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0102007>.
- LUO, Qiang et al. (2013). "Spatio-temporal Granger causality : A new framework". In : *NeuroImage* 79, p. 241–263.
- LUZEAUX, Dominique (2015). "A formal foundation of systems engineering". In : *Complex Systems Design & Management*. Springer, p. 133–148.
- MACHARIS, Cathy et al. (2010). "A decision analysis framework for intermodal transport : Comparing fuel price increases and the internalisation of external costs". In : *Transportation Research Part A : Policy and Practice* 44.7, p. 550–561.
- MAHMASSANI, Hani S et Gang-Len CHANG (1987). "On boundedly rational user equilibrium in transportation systems". In : *Transportation science* 21.2, p. 89–99.
- MAINZER, Klaus et Leon O CHUA (2013). *Local activity principle*. World Scientific.
- MAKSE, Hernán A et al. (1998). "Modeling urban growth patterns with correlated percolation". In : *Physical Review E* 58.6, p. 7054.
- MANGIN, David (2014). *Le Grand Paris, où en est-on ?* Conférence de David Mangin le 14 mars 2014, ENPC et ENSAVT.
- MANGIN, David et Philippe PANERAI (1999). *Projet urbain*. Parenthèses.
- MANSON, Steven M (2001). "Simplifying complexity : a review of complexity theory". In : *Geoforum* 32.3, p. 405–414.
- (2008). "Does scale exist? An epistemological scale continuum for complex human–environment systems". In : *Geoforum* 39.2, p. 776–788.
- MANTEGNA, Rosario Nunzio, Harry Eugene STANLEY et al. (2000). *An introduction to econophysics : correlations and complexity in finance*. T. 9. Cambridge university press Cambridge.
- MARCHIONNI, Caterina (2004). "Geographical economics versus economic geography : towards a clarification of the dispute". In : *Environment and Planning A* 36.10, p. 1737–1753.
- MARLER, R Timothy et Jasbir S ARORA (2004). "Survey of multi-objective optimization methods for engineering". In : *Structural and multidisciplinary optimization* 26.6, p. 369–395.
- MENDELEY (2015). *Mendeley Reference Manager*. <http://www.mendeley.com/>.
- MILLER, Harvey J (1999). "Measuring space-time accessibility benefits within transportation networks : basic theory and computational procedures". In : *Geographical analysis* 31.1, p. 1–26.
- MIMEUR, Christophe (2016). "The traces of speed between space and network". Theses. Université de Bourgogne Franche-Comté. URL : <https://hal.archives-ouvertes.fr/tel-01451164>.
- MIN, Wanli et Laura WYNTER (2011). "Real-time road traffic prediction with spatio-temporal correlations". In : *Transportation Research Part C : Emerging Technologies* 19.4, p. 606–616.

- MOECKEL, Rolf et al. (2003). "Creating a synthetic population". In : *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*.
- MONOD, J. (1970). *Le hasard et la Nécessité*. Points, Paris.
- MOORE, Christopher et Stephan MERTENS (2011). *The nature of computation*. OUP Oxford.
- MORENO REGAN, Omar (2016). "Etude du comportement des tunnels en maçonnerie du métro parisien". Thèse de doct. Paris Est.
- MORENO, Diego et al. (2012). "Un automate cellulaire pour expérimenter les effets de la proximité dans le processus d'étalement urbain : le modèle Raumulus". In : *Cybergeo : European Journal of Geography*.
- MOUDON, Anne Vernez (1997). "Urban morphology as an emerging interdisciplinary field". In : *Urban morphology* 1.1, p. 3–10.
- MOULIN-FRIER, C. et al. (2017). "Embodied Artificial Intelligence through Distributed Adaptive Control : An Integrated Framework". In : *ArXiv e-prints*. arXiv : 1704.01407 [cs.AI].
- MURPHY, Alexander B (2012). "Entente territorial : Sack and Raffestin on territoriality". In : *Environment and Planning D : Society and Space* 30.1, p. 159–172.
- NLTK (2015). *Natural Language Toolkit*, Stanford University.
- NEWMAN, M. E. J. (2013). "Prediction of highly cited papers". In : *ArXiv e-prints*. arXiv : 1310.8220 [physics.soc-ph].
- (2016). "Community detection in networks : Modularity optimization and maximum likelihood are equivalent". In : *ArXiv e-prints*. arXiv : 1606.02319.
- NEWMAN, MEJ (2011). "Complex systems : A survey". In : *arXiv preprint arXiv :1112.1440*.
- NEWMAN, Mark EJ (2003). "The structure and function of complex networks". In : *SIAM review* 45.2, p. 167–256.
- NICOSIA, Vincenzo et al. (2009). "Extending the definition of modularity to directed graphs with overlapping communities". In : *Journal of Statistical Mechanics : Theory and Experiment* 2009.03, Po3024.
- NIEDERREITER, H (1972). "Discrepancy and convex programming". In : *Annali di matematica pura ed applicata* 93.1, p. 89–97.
- NITSCH, Volker (2005). "Zipf zipped". In : *Journal of Urban Economics* 57.1, p. 86–100.
- OECD (2009). "OECD Patent Statistics Manual". In : doi : <http://dx.doi.org/10.1787/9789264056442-en>. URL : /content/book/9789264056442-en.
- O'SULLIVAN, David et Steven M MANSON (2015). "Do Physicists Have'Geography Envy'? And What Can Geographers Learn From It?" In : *Annals of the Association of American Geographers*.
- O'BRIEN, Oliver et al. (2014). "Mining bicycle sharing data for generating insights into sustainable transport systems". In : *Journal of Transport Geography* 34, p. 262–273.

- OFFNER, Jean-Marc (1993). "Les "effets structurants" du transport : mythe politique, mystification scientifique". In : *Espace géographique* 22.3, p. 233–242.
- OFFNER, Jean-Marc et Denise PUMAIN (1996). "Réseaux et territoires-significations croisées". In :
- OPENSTREETMAP (2012). *OpenStreetMap*.
- OPENSHAW, S (1983). *FROM DATA CRUNCHING TO MODEL CRUNCHING – THE DAWN OF A NEW ERA*.
- OPENSHAW, Stan (1984). *The Modifiable Areal Unit Problem*. Norwich, UK : Geo Books.
- OSTROWETSKY, S. & al. (2004). "Les Villes Nouvelles, 30 ans après". In : *Espaces et Sociétés* n°119, 4/2004.
- PARK, Inchae et Byungun YOON (2014). "A semantic analysis approach for identifying patent infringement based on a product-patent map". In : *Technology Analysis & Strategic Management* 26.8, p. 855–874.
- PAULLEY, Neil J et F Vernon WEBSTER (1991). "Overview of an international study to compare models and evaluate land-use and transport policies". In : *Transport Reviews* 11.3, p. 197–222.
- PAULUS, Fabien (2004). "Coévolution dans les systèmes de villes : croissance et spécialisation des aires urbaines françaises de 1950 à 2000". Thèse de doct. Université Panthéon-Sorbonne-Paris I.
- PEREZ-RIVEROL, Yasset et al. (2016). "Ten Simple Rules for Taking Advantage of Git and GitHub". In : *PLoS Comput Biol* 12.7, p. 1–11. DOI : [10.1371/journal.pcbi.1004947](https://doi.org/10.1371/journal.pcbi.1004947). URL : <http://dx.doi.org/10.1371%2Fjournal.pcbi.1004947>.
- PÉTONNET, Colette (1982). "L'Observation flottante L'exemple d'un cimetière parisien". In : *l'Homme*, p. 37–47.
- PFAENDER, Fabien (2009). "Spatialisation de l'information". Thèse de doct. Compiègne.
- PICON, Antoine (2013). *Smart cities : théorie et critique d'un idéal auto-réalisateur*. B2.
- PIGOZZI, Bruce Wm (1980). "Interurban linkages through polynomially constrained distributed lags". In : *Geographical Analysis* 12.4, p. 340–352.
- PIKETTY, Thomas (2013). *Le capital au XXIe siècle*. Seuil.
- PORTUGALI, Juval (2011). "SIRN–Synergetic Inter-Representation Networks". In : *Complexity, Cognition and the City*, p. 139–165.
- POTIRON, Y. (2016). "Estimating the integrated parameter of the locally parametric model in high-frequency data." In : *Working Paper*.
- POTIRON, Yoann et Per MYKLAND (2015). "Estimation of integrated quadratic covariation between two assets with endogenous sampling times". In : *arXiv preprint arXiv:1507.01033*.
- PRESCHITSCHKE, Nina et al. (2013). "Anticipating industry convergence : Semantic analyses vs IPC co-classification analyses of patents".

- In : *Foresight* 15.6. Sous la dir. de Tugrul DAIM, p. 446–464. ISSN : 1463-6689. DOI : [10.1108/fs-10-2012-0075](https://doi.org/10.1108/fs-10-2012-0075).
- PRIGOGINE, Ilya et Isabelle STENGERS (1997). *The end of certainty*. Simon et Schuster.
- PRITCHARD, David R et Eric J MILLER (2009). "Advances in agent population synthesis and application in an integrated land use and transportation model". In : *Transportation Research Board 88th Annual Meeting*. 09-1686.
- PUMAIN, Denise (1997). "Pour une théorie évolutive des villes". In : *Espace géographique* 26.2, p. 119–134.
- (2003). "Une approche de la complexité en géographie". In : *Géocarrefour* 78.1, p. 25–31.
 - (2005). "Cumulativité des connaissances". In : *Revue européenne des sciences sociales. European Journal of Social Sciences* XLIII-131, p. 5–12.
 - (2010). "Une théorie géographique des villes". In : *Bulletin de la Société géographie de Liège* 55, p. 5–15.
 - (2012a). "Multi-agent system modelling for urban systems : The series of SIMPOP models". In : *Agent-based models of geographical systems*. Springer, p. 721–738.
 - (2012b). "Urban systems dynamics, urban growth and scaling laws : The question of ergodicity". In : *Complexity Theories of Cities Have Come of Age*. Springer, p. 91–103.
- PUMAIN, Denise et al. (2009). "Innovation cycles and urban dynamics". In : *Complexity perspectives in innovation and social change*, p. 237–260.
- PUMAIN, Denise et Romain REUILLOU (2017). *Urban Dynamics and Simulation Models*.
- PUMAIN, Denise et al. (2006). "An evolutionary theory for interpreting urban scaling laws". In : *Cybergeo : European Journal of Geography*.
- PUTMAN, Stephen H (1975). "Urban land use and transportation models : A state-of-the-art summary". In : *Transportation Research* 9.2, p. 187–202.
- PUZIS, Rami et al. (2013). "Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks". In : *Journal of Intelligent Transportation Systems* 17.1, p. 91–105.
- QGIS, DT (2011). "Quantum GIS geographic information system". In : *Open Source Geospatial Foundation Project*.
- R CORE TEAM (2015a). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL : <http://www.R-project.org/>.
- (2015b). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL : <http://www.R-project.org/>.

- RADDICK, M. J. et al. (2010). "Galaxy Zoo : Exploring the Motivations of Citizen Science Volunteers". In : *Astronomy Education Review* 9.1, p. 010103. DOI : [10 . 3847 / AER2009036](https://doi.org/10.3847/AER2009036). arXiv : [0909 . 2925 \[astro-ph.IM\]](https://arxiv.org/abs/0909.2925).
- RAFFESTIN, Claude (1988). "Repères pour une théorie de la territorialité humaine". In :
- RAIMBAULT, J. (2015a). "Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach". In : *Forthcoming in ECTQG 2015 Proceedings*.
- (2016a). "Generation of Correlated Synthetic Data". In : *Forthcoming in Actes des Journées de Rochebrune 2016*.
- RAIMBAULT, J. et al. (2014). "A hybrid network/grid model of urban morphogenesis and optimization". In : *Proceedings of the 4th International Conference on Complex Systems and Applications (ICCSA 2014), June 23-26, 2014, Université de Normandie, Le Havre, France ; M. A. Aziz-Alaoui, C. Bertelle, X. Z. Liu, D. Olivier, eds. : pp. 51-60.*
- RAIMBAULT, J. et J. GONZALEZ (May 2015). "Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport". In : *Rencontres du Labex Dynamites*.
- RAIMBAULT, Juste (2015b). "Hybrid Modeling of a Bike-Sharing Transportation System". In : *International Conference on Computational Social Science*.
- (2016b). "A Discrepancy-based Framework to Compare Robustness between Multi-Attribute Evaluations". In : *Forthcoming in Proceedings of CSDM 2016. arXiv preprint arXiv :1608.00840*.
- (2016c). "For a Cautious Use of Big Data and Computation". In : *Royal Geographical Society-Annual Conference 2016-Session : Geocomputation, the Next 20 Years (1)*.
- (2016d). "Indirect Bibliometrics by Complex Network Analysis". In : *20e Anniversaire de Cybergeo*.
- (2016e). "Investigating the Empirical Existence of Static User Equilibrium". In : *Forthcoming in Transportation Research Procedia, EWGT2016. arXiv preprint arXiv :1608.05266*.
- (2016f). "Models of growth for system of cities : Back to the simple". In : *Conference on Complex Systems 2016*.
- (2016g). "Towards Models Coupling Urban Growth and Transportation Network Growth. First year preliminary memoire. DOI : <http://dx.doi.org/10.5281/zenodo.60538>". Thèse de doct. Université Paris-Diderot - Paris VII.
- (2017a). "An Applied Knowledge Framework to Study Complex Systems". In : *Forthcoming in CSDM2017 proceedings*.
- (2017b). "Co-construire Modèles, Etudes Empiriques et Théories en Géographie Théorique et Quantitative : le cas des Interactions entre Réseaux et Territoires". In : *Treizièmes Rencontres de Théo-Quant*.

- RAIMBAULT, Juste (2017c). "Complexity, Complexities and Complex Knowledges". In : *forthcoming discussion at ERC Divercity Workshop, 12-13 October 2017*.
- (2017d). *Entretiens vo.2 [Data set]*. Zenodo. <http://doi.org/10.5281/zenodo.556331>.
 - (2017e). "Identification de Causalités dans des Données Spatio-temporelles". In : *submitted to SAGEO 2017*.
 - (2017f). "Investigating the Empirical Existence of Static User Equilibrium". In : *Transportation Research Procedia* 22C, p. 450–458. URL : DOI : [10.1016/j.trpro.2017.03.053](https://doi.org/10.1016/j.trpro.2017.03.053); arXiv preprint arXiv:1608.05266.
 - (2017g). "Un Cadre de Connaissances pour une Géographie Intégrée". In : *Journée des jeunes chercheurs de l'Institut de Géographie de Paris*. Paris, France. URL : <https://halshs.archives-ouvertes.fr/halshs-01505084>.
- RAIMBAULT, Juste et Antonin BERGEAUD (2017). "The Cost of Transportation : Spatial Analysis of Fuel Prices in the US". In : *Forthcoming in Transportation Research Procedia, EWGT2017*.
- RAM, Karthik (2013). "Git can facilitate greater reproducibility and increased transparency in science." In : *Source code for biology and medicine* 8.1, p. 7.
- RAMSEY, James B (2002). "Wavelets in economics and finance : Past and future". In : *Studies in Nonlinear Dynamics & Econometrics* 6.
- RASMUSSEN, Thomas Kjær et al. (2015). "Stochastic user equilibrium with equilibrated choice sets : Part II—Solving the restricted SUE for the logit family". In : *Transportation Research Part B : Methodological* 77, p. 146–165.
- READ, Dwight et al. (2009). "The innovation innovation". In : *Complexity perspectives in innovation and social change*. Springer, p. 43–84.
- REID, Chris R et al. (2016). "Decision-making without a brain : how an amoeboid organism solves the two-armed bandit". In : *Journal of The Royal Society Interface* 13.119, p. 20160030.
- REUILLOUN, Romain et al. (2013). "OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models". In : *Future Generation Computer Systems* 29.8, p. 1981–1990.
- REUILLOUN, Romain et al. (2015). "A New Method to Evaluate Simulation Models : The Calibration Profile (CP) Algorithm". In : *Journal of Artificial Societies and Social Simulation* 18.1, p. 12. ISSN : 1460-7425. URL : <http://jasss.soc.surrey.ac.uk/18/1/12.html>.
- REY-COYREHOURCQ, Sébastien (2015). "Une plateforme intégrée pour la construction et Une plateforme intégrée pour la construction et l'évaluation de modèles de simulation en géographie". Thèse de doct. Université Paris 1 Panthéon-Sorbonne.
- REYMOND, Henri et Colette CAUVIN (2013). "La logique ternaire de Stéphane Lupasco et le raisonnement géocartographique bioculturel d'*Homo geographicus*. L'apport de la notion de couplage trans-

- disciplinaire dans l'approche de l'agrégation morphologique des agglomérations urbaines". In : *Cybergeo : European Journal of Geography*.
- RIETVELD, Piet et al. (2001). "Spatial graduation of fuel taxes ; consequences for cross-border and domestic fuelling". In : *Transportation Research Part A : Policy and Practice* 35.5, p. 433–457.
- RIETVELD, Piet et Stefan van WOUDENBERG (2005). "Why fuel prices differ". In : *Energy Economics* 27.1, p. 79–92.
- RIPOLL, Fabrice (2017). "Géographie de l'alternatif, Géographies alternatives ? Grand Témoin." In : *Journée des Jeunes Chercheurs de l'Institut de Géographie*.
- ROMER, Paul M (1990). "Endogenous Technological Change". In : *Journal of Political Economy* 98.5, S71–102. URL : <https://ideas.repec.org/a/ucp/jpolec/v98y1990i5ps71-102.html>.
- ROTH, Camille (2009). "Reconstruction Failures : Questioning Level Design". In : *Epistemological Aspects of Computer Simulation in the Social Sciences*. Springer, p. 89–98.
- RUBNER, Yossi et al. (2000). "The earth mover's distance as a metric for image retrieval". In : *International journal of computer vision* 40.2, p. 99–121.
- RUCKER, Gerta (2012). "Network meta-analysis, electrical networks and graph theory". In : *Research Synthesis Methods* 3.4, p. 312–324.
- RUI, Yikang et Yifang BAN (2011). "Urban growth modeling with road network expansion and land use development". In : *Advances in Cartography and GIScience. Volume 2*. Springer, p. 399–412.
- RUI, Yikang et al. (2013). "Exploring the patterns and evolution of self-organized urban street networks through modeling". In : *The European Physical Journal B* 86.3, p. 1–8.
- SDRIF (2013). *Île-de-France 2030. ORIENTATIONS RÉGLEMENTAIRES ET CARTE DE DESTINATION GÉNÉRALE DES DIFFÉRENTES PARTIES DU TERRITOIRE*.
- STIF (2010). *ArcExpress, débat public sur le métro de rocade. Dossier du Maître d'Ouvrage*. archived at http://archive.wikiwix.com/cache/?url=http%3A%2F%2Fwww.debatpublic-arcxpress.org%2F_script%2Fntsp-document-file_download.php%3Fdocument_id%3D92%26document_file_id%3D106.
- SAMANIEGO, Horacio et Melanie E MOSES (2008). "Cities as organisms : Allometric scaling of urban road networks". In : *Journal of Transport and Land use* 1.1.
- SANDERS, Lena (1992). *Système de villes et synergétique*. Economica.
- SANDERS, Lena et al. (1997). "SIMPOP : a multiagent system for the study of urbanism". In : *Environment and Planning B* 24, p. 287–306.
- SARIGÖL, E. et al. (2014). "Predicting Scientific Success Based on Co-authorship Networks". In : *ArXiv e-prints*. arXiv : 1402.7268 [physics.soc-ph].

- SCHMITT, Clara (2014). "Modélisation de la dynamique des systèmes de peuplement : de SimpopLocal à SimpopNet." Thèse de doct. Paris 1.
- SCHMITT, Clara et al. (2014). "Half a billion simulations : Evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model". In :
- SHALIZI, Cosma Rohilla et James P CRUTCHFIELD (2001). "Computational mechanics : Pattern and prediction, structure and simplicity". In : *Journal of statistical physics* 104.3-4, p. 817–879.
- SIMON, Herbert A. (1955). "On a Class of Skew Distribution Functions". English. In : *Biometrika* 42.3 / 4, pp. 425–440. ISSN : 00063444. URL : <http://www.jstor.org/stable/2333389>.
- SORENSEN, Olav et al. (2006). "Complexity, networks and knowledge flow". In : *Research policy* 35.7, p. 994–1017.
- SOUAMI, Taoufik (2012). *Ecoquartiers : secrets de fabrication*. Scrineo.
- STANLEY, H Eugene et al. (1999). "Econophysics : Can physicists contribute to the science of economics ?" In : *Physica A : Statistical Mechanics and its Applications* 269.1, p. 156–169.
- STEVENS, Forrest R. et al. (2015). "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data". In : *PLoS ONE* 10.2, p. 1–22. DOI : [10.1371/journal.pone.0107042](https://doi.org/10.1371/journal.pone.0107042). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0107042>.
- STODDEN, Victoria (2010). "The scientific method in practice : Reproducibility in the computational sciences". In :
- SULLIVAN, JL et al. (2010). "Identifying critical road segments and measuring system-wide robustness in transportation networks with isolating links : a link-based capacity-reduction approach". In : *Transportation Research Part A : Policy and Practice* 44.5, p. 323–336.
- SWERTS, Elfie et Eric DENIS (2015). "Megacities : The Asian Era". In : *Urban Development Challenges, Risks and Resilience in Asian Mega Cities*. Springer, p. 1–28.
- TAN, Wei et al. (2013). "Social-network-sourced big data analytics". In : *IEEE Internet Computing* 17.5, p. 62–69.
- TARDY, Christine (2004). "The role of English in scientific communication : lingua franca or Tyrannosaurus rex ?" In : *Journal of English for academic purposes* 3.3, p. 247–269.
- TEAM, R Core (2000). *R Language Definition*.
- TEPLITSKY, M. et al. (2015). "Amplifying the Impact of Open Access : Wikipedia and the Diffusion of Science". In : *ArXiv e-prints*. arXiv : [1506.07608 \[cs.DL\]](https://arxiv.org/abs/1506.07608).
- TERO, Atsushi et al. (2006). "Physarum solver : a biologically inspired method of road-network navigation". In : *Physica A : Statistical Mechanics and its Applications* 363.1, p. 115–119.

- TERO, Atsushi et al. (2010). "Rules for Biologically Inspired Adaptive Network Design". In : *Science* 327.5964, p. 439–442. DOI : [10.1126/science.1177894](https://doi.org/10.1126/science.1177894). eprint : <http://www.sciencemag.org/content/327/5964/439.full.pdf>. URL : <http://www.sciencemag.org/content/327/5964/439.abstract>.
- THOM, René (1974). "Stabilité structurelle et morphogénèse". In : *Poetics* 3.2, p. 7–19.
- THOMAS, Isabelle et al. (2017). "City delineation in European applications of LUTI models : review and tests". In : *Transport Reviews* 0.0, p. 1–27. DOI : [10.1080/01441647.2017.1295112](https://doi.org/10.1080/01441647.2017.1295112). eprint : <http://www.tandfonline.com/doi/pdf/10.1080/01441647.2017.1295112>. URL : <http://www.tandfonline.com/doi/abs/10.1080/01441647.2017.1295112>.
- TIVADAR, Mihai et al. (2014). "OASIS—un Outil d'Analyse de la Ségrégation et des Inégalités Spatiales". In : *Cybergeo : European Journal of Geography*.
- TORDEUX, Antoine et Sylvain LASSARRE (2016). "Jam avoidance with autonomous systems". In : *arXiv preprint arXiv:1601.07713*.
- TSAI, Yu-Hsin (2005). "Quantifying urban form : compactness versus' sprawl''. In : *Urban studies* 42.1, p. 141–161.
- TSAY, Ruey S. (2015). *MTS : All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 0.33. URL : <http://CRAN.R-project.org/package=MTS>.
- TSENG, Yuen-Hsien et al. (2007). "Text mining techniques for patent analysis". In : *Information Processing & Management* 43.5, p. 1216–1247.
- TUMMINELLO, Michele et al. (2005). "A tool for filtering information in complex systems". In : *Proceedings of the National Academy of Sciences of the United States of America* 102, p. 10421–10426.
- TURING, Alan Mathison (1952). "The chemical basis of morphogenesis". In : *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 237.641, p. 37–72.
- VALLES-CATALA, Toni et al. (2016). "Multilayer stochastic block models reveal the multilayer structure of complex networks". In : *Physical Review X* 6.1, p. 011036.
- VARENNE, Franck (2010a). "Framework for M&S with Agents in Regard to Agent Simulations in Social Sciences". In : *Activity-Based Modeling and Simulation*, p. 53–84.
- (2010b). "Les simulations computationnelles dans les sciences sociales". In : *Nouvelles Perspectives en Sciences Sociales* 5.2, p. 17–49.
- VARENNE, Franck, Marc SILBERSTEIN et al. (2013). *Modéliser & simuler. Epistémologies et pratiques de la modélisation et de la simulation, tome 1*.

- VARET, Suzanne (2010). "Développement de méthodes statistiques pour la prédiction d'un gabarit de signature infrarouge". Thèse de doct. Université Paul Sabatier-Toulouse III.
- VATTAY, Gabor et al. (2015). "Quantum Criticality at the Origin of Life". In : *arXiv preprint arXiv :1502.06880*.
- VEIGA, Allan Koch et al. (2017). "A conceptual framework for quality assessment and management of biodiversity data". In : *PLOS ONE* 12.6, p. 1–20. DOI : [10.1371/journal.pone.0178731](https://doi.org/10.1371/journal.pone.0178731). URL : <https://doi.org/10.1371/journal.pone.0178731>.
- VERLINDE, E. P. (2016). "Emergent Gravity and the Dark Universe". In : *ArXiv e-prints*. arXiv : [1611.02269 \[hep-th\]](https://arxiv.org/abs/1611.02269).
- VISSEER, Hans et T DE NIJS (2006). "The map comparison kit". In : *Environmental Modelling & Software* 21.3, p. 346–358.
- WANG, Jiang-Jiang et al. (2009). "Review on multi-criteria decision analysis aid in sustainable energy decision-making". In : *Renewable and Sustainable Energy Reviews* 13.9, p. 2263–2278.
- WANG, Y.-S. et al. (2017). "Separable and Localized System Level Synthesis for Large-Scale Systems". In : *ArXiv e-prints*. arXiv : [1701.05880 \[math.OC\]](https://arxiv.org/abs/1701.05880).
- WARDROP, John Glen (1952). "Some theoretical aspects of road traffic research." In : *Proceedings of the institution of civil engineers* 1.3, p. 325–362.
- WATSON, Benjamin et al. (2008). "Procedural urban modeling in practice". In : *IEEE Computer Graphics and Applications* 3, p. 18–26.
- WEGENER, Michael et Franz FÜRST (2004). "Land-use transport interaction : state of the art". In : *Available at SSRN* 1434678.
- WEGENER, Michael et al. (1991). "One city, three models : comparison of land-use/transport policy simulation models for Dortmund". In : *Transport Reviews* 11.2, p. 107–129.
- WEIBULL, Jörgen W (1976). "An axiomatic approach to the measurement of accessibility". In : *Regional Science and Urban Economics* 6.4, p. 357–379.
- WHITEHAND, JWR et al. (1999). "Urban morphogenesis at the microscale : how houses change". In : *Environment and Planning B : Planning and Design* 26.4, p. 503–515.
- WIENER, Norbert (1948). *Cybernetics*. Hermann Paris.
- WILENSKY, Uri (1999). "NetLogo". In :
- WILSON, A. (1981). *Catastrophe theory and bifurcation : Application to Urban and Regional System*. London : Croom Helm.
- WILSON, G et al. (2017). "Good enough practices in scientific computing". In : *PLoS Comput Biol* 13.6, e1005510.
- WOLFRAM, Stephen (2002). *A new kind of science*. T. 5. Wolfram media Champaign.
- XIE, Feng et David LEVINSON (2009a). "How streetcars shaped suburbanization : a Granger causality analysis of land use and transit in the Twin Cities". In : *Journal of Economic Geography*, lbp031.

- (2009b). "Jurisdictional control and network growth". In : *Networks and Spatial Economics* 9.3, p. 459–483.
- (2009c). "Modeling the growth of transportation networks : A comprehensive review". In : *Networks and Spatial Economics* 9.3, p. 291–307.
- XIE, Yihui (2013). "knitr : A general-purpose package for dynamic report generation in R". In : *R package version 1.7*.
- YAMASAKI, Kazuko et al. (2006). "Preferential attachment and growth dynamics in complex systems". In : *Physical Review E* 74.3, p. 035103.
- YAMINS, Daniel et al. (2003). "Growing urban roads". In : *Networks and Spatial Economics* 3.1, p. 69–85.
- YANG, Yiming et al. (2000). "Improving text categorization methods for event tracking". In : *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 65–72.
- YE, Xin (2011). "Investigation of Underlying Distributional Assumption in Nested Logit Model Using Copula-Based Simulation and Numerical Approximation". In : *Transportation Research Record : Journal of the Transportation Research Board* 2254, p. 36–43.
- YERRA, Bhanu M et David M LEVINSON (2005). "The emergence of hierarchy in transportation networks". In : *The Annals of Regional Science* 39.3, p. 541–553.
- YOON, Byungun et Yongtae PARK (2004). "A text-mining-based patent network : Analytical tool for high-technology trend". In : *The Journal of High Technology Management Research* 15.1, p. 37–50.
- YOON, Janghyeok et Kwangsoo KIM (2011). "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection". In : *Scientometrics* 90.2, p. 445–461.
- YOUN, Hyejin et al. (2015). "Invention as a combinatorial process : evidence from US patents". In : *Journal of The Royal Society Interface* 12.106. ISSN : 1742-5689. DOI : [10.1098/rsif.2015.0272](https://doi.org/10.1098/rsif.2015.0272).
- ZEMBRI, Pierre (1997). "Les fondements de la remise en cause du Schéma Directeur des liaisons ferroviaires à grande vitesse : des faiblesses avant tout structurelles". In : *Annales de géographie*. JSTOR, p. 183–194.
- (2008). "La contribution de la grande vitesse ferroviaire à l'interrégionalité en France.(High-speed rail and inter-regionality in France)". In : *Bulletin de l'Association de géographes français* 85.4, p. 443–460.
- (2010). "The new purposes of the French high-speed rail system in the framework of a centralized network : a substitute to the domestic air transport market?" In :
- ZHANG, Kuilin et al. (2013). "Dynamic pricing, heterogeneous users and perception error : Probit-based bi-criterion dynamic stochastic user equilibrium assignment". In : *Transportation Research Part C : Emerging Technologies* 27, p. 189–204.

- ZHANG, Lei et David LEVINSON (2007). "The economics of transportation network growth". In : *Essays on transport economics*. Springer, p. 317–339.
- ZHU, Liping et al. (2013a). "Amoeba-based computing for traveling salesman problem : Long-term correlations between spatially separated individual cells of *Physarum polycephalum*". In : *Biosystems* 112.1, p. 1–10.
- ZHU, Shanjiang et David LEVINSON (2010). "Do people use the shortest path ? An empirical test of Wardrop's first principle". In : *91th annual meeting of the Transportation Research Board, Washington*. T. 8. Citeseer.
- ZHU, Yaojia et al. (2013b). "Scalable text and link analysis with mixed-topic link models". In : *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 473–481.
- ZILSEL (2015). "LA POSITION DE LA REVUE SOCIÉTÉS DANS L'ESPACE DISCURSIF DE LA SOCIOLOGIE FRANÇAISE". In : <http://zilsel.hypotheses.org/category/canular>.
- ZIMAN, John (2003). *Technological innovation as an evolutionary process*. Cambridge University Press.
- TEAM, Osmosis (2016). OSMOSIS. <http://wiki.openstreetmap.org/wiki/Osmosis>.

Quatrième partie

APPENDICES

A

INFORMATIONS SUPPLÉMENTAIRES

C'est hardcore tes calculs.

- ANONYME

This chapter gathers various technical developments, that have the common points to be not essential to the core of the thesis and difficult to digest.

A.1 ELEMENTS DE TERRAIN

A.1.1 Carnet de Terrain

Nous rendons compte ici avec un certain niveau de détail des différentes sorties de terrain alimentant la section [1.3](#).

A.2 TECHNICAL DEVELOPMENTS

A.2.1 Dérivations pour les modèles de croissance urbaine

Lemma 1 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

Proof Starting with first moment, we denote $\bar{P}_i(t) = \mathbb{E}[P_i(t)]$. Independence of Gibrat growth rate yields directly $\bar{P}_i(t) = \mathbb{E}[R_i(t)] \cdot \bar{P}_i(t-1)$. Starting for the preferential attachment model, we have $\bar{P}_i(t) = \mathbb{E}[P_i(t)] = \sum_{k=0}^{+\infty} k \mathbb{P}[P_i(t) = k]$. But

$$\{P_i(t) = k\} = \bigcup_{\delta=0}^{\infty} (\{P_i(t-1) = k-\delta\} \cap \{P_i \leftarrow P_i + 1\}^{\delta})$$

where the second event corresponds to city i being increased δ times between $t-1$ and t (note that events are empty for $\delta \geq k$). Thus, being careful on the conditional nature of preferential attachment formulation, stating that $\mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] = \lambda \cdot \frac{p}{\bar{P}(t-1)}$ (total population $P(t)$ assumed deterministic), we obtain

$$\begin{aligned} \mathbb{P}[\{P_i \leftarrow P_i + 1\}] &= \sum_p \mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] \cdot \mathbb{P}[P_i(t-1) = p] \\ &= \sum_p \lambda \cdot \frac{p}{\bar{P}(t-1)} \mathbb{P}[P_i(t-1) = p] = \lambda \cdot \frac{\bar{P}_i(t-1)}{\bar{P}(t-1)} \end{aligned}$$

It gives therefore, knowing that $P(t-1) = P_0 + m \cdot (t-1)$ and denoting $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)}$

$$\begin{aligned}\bar{P}_i(t) &= \sum_{k=0}^{\infty} \sum_{\delta=0}^{\infty} k \cdot \left(\lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \right)^{\delta} \cdot \mathbb{P}[P_i(t-1) = k - \delta] \\ &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta') \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\ &= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot (\delta' + \bar{P}_i(t-1)) = \frac{q}{(1-q)^2} + \frac{\bar{P}_i(t-1)}{(1-q)} \\ &= \frac{\bar{P}_i(t-1)}{1-q} \left[1 + \frac{1}{\bar{P}_i(t-1)} \frac{q}{(1-q)} \right]\end{aligned}$$

As it is not expected to have $\bar{P}_i(t) \ll P(t)$ (fat tail distributions), a limit can be taken only through λ . Taking $\lambda \ll 1$ yields, as $0 < \bar{P}_i(t)/P(t) < 1$, that $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \ll 1$ and thus we can expand in first order of q , what gives $\bar{P}_i(t) = \bar{P}_i(t-1) \cdot \left[1 + \left(1 + \frac{1}{\bar{P}_i(t-1)} \right) q + o(q) \right]$

$$\bar{P}_i(t) \simeq \left[1 + \frac{\lambda}{P_0 + m \cdot (t-1)} \right] \cdot \bar{P}_i(t-1)$$

It means that this limit is equivalent in expectancy to a Gibrat model with $\mu_i(t) = \mu(t) = 1 + \frac{\lambda}{P_0 + m \cdot (t-1)}$.

For the second moment, we can do an analog computation. We have still

$$\mathbb{E}[P_i(t)^2] = \mathbb{E}[R_i(t)^2] \cdot \mathbb{E}[P_i(t-1)^2]$$

and

$$\mathbb{E}[P_i(t)^2] = \sum_{k=0}^{+\infty} k^2 \mathbb{P}[P_i(t) = k]$$

We obtain the same way

$$\begin{aligned}\mathbb{E}[P_i(t)^2] &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta')^2 \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\ &= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot \left(\mathbb{E}[P_i(t-1)^2] + 2\delta' \bar{P}_i(t-1) + \delta'^2 \right) \\ &= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} + \frac{2q \bar{P}_i(t-1)}{(1-q)^2} + \frac{q(q+1)}{(1-q)^3} \\ &= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} \left[1 + \frac{q}{\mathbb{E}[P_i(t-1)^2]} \left(\frac{2\bar{P}_i(t-1)}{1-q} + \frac{(1+q)}{(1-q)^2} \right) \right]\end{aligned}$$

We have therefore an equivalence between the Gibrat model as a continuous formulation of a Preferential Attachment (or Simon model) in a certain limit. ■

A.2.2 Sensibilité des Lois d'Echelle Urbaines

We formalize the simple theoretical context in which we will derive the sensitivity of scaling to city definition. Let consider a polycentric city system, which spatial density distributions can be reasonably constructed as the superposition of monocentric fast-decreasing spatial kernels, such as an exponential mixture model [ANAS, ARNOTT et SMALL, 1998]. Taking a geographical space as \mathbb{R}^2 , we take for any $\vec{x} \in \mathbb{R}^2$ **C : (Florent) attention à la sensibilité de certains géographes** the density of population as

$$d(\vec{x}) = \sum_{i=1}^N d_i(\vec{x}) = \sum_{i=1}^N d_i^0 \cdot \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{r_i}\right) \quad (11)$$

where r_i are spread parameters of kernels, d_i^0 densities at origins, \vec{x}_i positions of centers. We furthermore assume the following constraints :

1. To simplify, cities are monocentric, in the sense that for all $i \neq j$, we have $\|\vec{x}_i - \vec{x}_j\| \gg r_i$.
2. It allows to impose structural scaling in the urban system by the simple constraint on city populations P_i . One can compute by integration that $P_i = 2\pi d_i^0 r_i^2$, what gives by injection into the scaling hypothesis $\ln P_i = \ln P_{\max} - \alpha \ln i$, the following relation between parameters : $\ln [d_i^0 r_i^2] = K' - \alpha \ln i$.

To study scaling relations, we consider a random scalar spatial variable $a(\vec{x})$ representing one aspect of the city, that can be everything but has the dimension of a spatial density, such that the indicator $A(D) = \mathbb{E}[\iint_D a(\vec{x}) d\vec{x}]$ represents the expected quantity of a in area D . We make the assumption that $a \in \{0; 1\}$ ("counting" indicator) and that its law is given by $P[a(\vec{x}) = 1] = f(d(\vec{x}))$. Following the empirical work done in [COTTINEAU, 2015], the integrated indicator on city i as a function of θ is given by

$$A_i(\theta) = A(D(\vec{x}_i, \theta))$$

where $D(\vec{x}_i, \theta)$ is the area centered in \vec{x}_i where $d(\vec{x}) > \theta$. Assumption 1 ensures that the area are roughly disjoint circles. We take furthermore a simple amenity such that it follows a local scaling law in the sense that $f(d) = \lambda \cdot d^\beta$. It seems a reasonable assumption since it was shown that many urban variable follow a fractal behavior at the intra-urban scale [KEERSMAECKER, FRANKHAUSER et THOMAS, 2003] and that it implies necessarily a power-law distribution [CHEN, 2010]. We make the additional assumption that $r_i = r_0$ does not depend on i , what is reasonable if the urban system is considered from a large scale. This assumption should be relaxed in numerical simulations. The estimated scaling exponent $\alpha(\theta)$ is then the result of the log-regression of $(A_i(\theta))_i$ against $(P_i(\theta))_i$ where $P_i(\theta) = \iint_{D(\vec{x}_i, \theta)} d$.

A.2.3 Dérivation Analytique de la Sensibilité

With above notations, let derive the expression of estimated exponent for quantity a as a function of density threshold parameter θ . The quantity computed for a given city i is, thanks to the monocentric assumption and in a spatial range and a range for θ such that $\theta \gg \sum_{j \neq i} d_j(\vec{x})$, allowing to approximate $d(\vec{x}) \simeq d_i(\vec{x})$ on $D(\vec{x}_i, \theta)$, is computed by

$$\begin{aligned} A_i(\theta) &= \lambda \cdot \iint_{D(\vec{x}_i, \theta)} d^\beta = 2\pi\lambda d_i^0 \beta \int_{r=0}^{r_0 \ln \frac{d_i^0}{\theta}} r \exp\left(-\frac{r\beta}{r_0}\right) dr \\ &= \frac{2\pi d_i^0 \beta r_0^2}{\beta^2} \left[1 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta - \left(\frac{\theta}{d_i^0} \right)^\beta \right] \end{aligned}$$

We obtain in a similar way the expression of $P_i(\theta)$

$$P_i(\theta) = 2\pi d_i^0 r_0^2 \left[1 + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0} - \frac{\theta}{d_i^0} \right]$$

The Ordinary-Least-Square estimation, solving the problem $\inf_{\alpha, C} \|(\ln A_i(\theta) - C - \alpha \ln P_i(\theta))_i\|^2$, gives the value $\alpha(\theta) = \frac{\text{Cov}[(\ln A_i(\theta))_i, (\ln P_i(\theta))_i]}{\text{Var}[(\ln P_i(\theta))_i]}$. As we work on city boundaries, threshold is expected to be significantly smaller than center density, i.e. $\theta/d_i^0 \ll 1$. We can develop the expression in the first order of θ/d_i^0 and use the global scaling law for city sizes, what gives $\ln A_i(\theta) \simeq K_A - \alpha \ln i + (\beta - 1) \ln d_i^0 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta$ and $\ln P_i(\theta) = K_P - \alpha \ln i + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0}$. Developing the covariance and variance gives finally an expression of the scaling exponent as a function of θ , where k_j, k'_j are constants obtained in the development :

$$\alpha(\theta) = \frac{k_0 + k_1 \theta + k_2 \theta^\beta + k_3 \theta^{\beta+1} + k_4 \theta \ln \theta + k_5 \theta^\beta \ln \theta + k_6 \theta^\beta (\ln \theta)^2 + k_7 \theta^{\beta+1} (\ln \theta)^2}{k'_0 + k'_1 \ln \theta + k'_2 \theta \ln \theta + k'_3 \theta^2 + k'_4 \theta^2 \ln \theta + k'_5 \theta^2 (\ln \theta)^2} \quad (12)$$

This rational fraction predicts the evolution of the scaling exponent when the threshold varies. We study numerically its behavior in the next section, among other numerical experiments.

A.2.4 Simulations Numériques

IMPLÉMENTATION **C : (Florent) définir ton champ d'investigation (des grilles carrées de taille prédéfinies, ce n'est pas du tout standard)**

We implement empirically the density model given in section A.2.2. Centers are successively chosen such that in a given region of space only one kernel dominates in the sense that the sum of other contributions are above a given threshold θ_e . **C : (Florent) est-ce toujours**

FIGURE 28 :

possible, y'a t-il unicité du centre ? Par quelle méthode précise détermine tu le centre ? In practice, adapting N to world size allows to respect the monocentric condition. Population are distributed in order to follow the scaling law with fixed α and r_i (arbitrary choice) by computing corresponding d_i^0 . Technical details of the implementation done in R [R CORE TEAM, 2015b] and using the package kernlab for efficient kernel mixture methods [KARATZOGLOU et al., 2004] are given as comments in source code¹. **C : (Florent) cela ne suffit pas, il faut en dire plus sur la méthode A1** : sure, surtout qu'on formule cette requete dans la partie méthodologique précédente, tout cela est un peu contradictoire.. We show in figure 28 example of synthetic density distributions on which the numerical study is conducted. The validation of theoretical results on these experimental mixtures must still be conducted, along with sensitivity tests to random perturbations, influence of kernel type, and two-parameters phase diagram when adding in the computational model functional density distribution and associated cut-off threshold.

C : (Florent) TB mais la encore, on ne sait pas précisément pourquoi tu te lances là dedans

PERTURBATIONS ALÉATOIRES The simple model used is quite reducing for maximal densities and radius distribution. We aim to proceed to an empirical study of the influence of noise in the system by fixing d_i^0 and r_i the following way :

- d_i^0 follows a reversed log-normal distribution with maximal value being a realistic maximal density
- Radiiuses are computed to respect rank-size law and then perturbed by a white noise. **C : (Florent) pourquoi ?**

TYPE DE NOYAU We shall test the influence of the type of spatial kernel used on results. We can test gaussian kernels and quadratic kernels with parameters within reasonable ranges analog to the exponential kernel.

A.3 EXPLORATION DES MODÈLES

This appendix gathers more precise model explorations, generally needed to support conclusions in main text but too long or repetitive to be included.

¹ available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Scaling>

FIGURE 29 :

A.4 CAUSALITÉS DANS LE MODÈLE RBD

B

METHODOLOGICAL DEVELOPMENTS

Développements Méthodologiques

We are now building a rigorous Science of Cities, contrarily to what was done before.

- MARC BARTHÉLÉMY C : (Arnaud) référence? A1 : EMCSS Fall 2014, Network Course Introduction

Such a shocking phrase **C : (Florent) je crois que si tu t'appuies explicitement sur la mise en exergue alors ce n'est plus une mise en exergue** was pronounced during the introduction of a *Network* course for students of Complex System Science. Besides the fact that the spirit of CSS **C : (Florent) pas mettre trop d'acronymes que tu ne réutiliseras pas** is precisely the opposite, i. e. the construction of integrative disciplines (vertical integration that is necessarily founded on the existing body of knowledge of concerned fields) that answer transversal questions (horizontal integration that imply interdisciplinarity) - see e. g. the roadmap for CS [BOURGINE, CHAVALARIS et AL., 2009], it reveals how methodological considerations shape the perceptions of disciplines. From a background in Physics, **C : (Florent) soit on connaît ton background?** “rigorous” implies the use of tools and methods judged more rigorous (analytical derivations, large datasets statistics, etc.). **C : (Florent) je ne suis pas sûr que cela soit ça la rigueur physicienne. ce serait plutôt un raisonnement sans trou du début à la fin sur des objets clairement définis; en sciences sociales il y a fréquemment des trous** But what is rigorous for someone will not be for an other discipline¹, depending on the purpose of each piece of research (perspectivism [GIERE, 2010c] poses the *model*, that includes methods, as the articulating core of research enterprises). Thus the full role of methodology aside and not beside theory and experiments. We go in this chapter into various methodological developments which may be precisely used later or contribute to the global background.

¹ a funny but sad anecdote told by a friend comes to mind : defending his PhD in statistics, he was told at the end by economists how they were impressed by the mathematical rigor of his work, whereas a mathematician judged that “he could have done everything on the back of an enveloppe”. **C : (Florent) ce n'est pas lié à la rigueur**

B.1 UN CADRE UNIFIÉ POUR LES MODÈLES STOCHASTIQUES DE CROISSANCE URBAINE

Urban growth modeling fall in the case of tentatives to find self-consistent rules reproducing dynamics of an urban system, and thus in our logic of system morphogenesis. **C : (Florent) est ce que faire de la morphogenèse est le but ou le moyen? ce n'est pas clair en lisant** We examine here methodological issues linked to different frameworks of urban growth.

B.1.1 *Introduction*

Various stochastic models aiming to reproduce population patterns on large temporal and spatial scales (city systems) have been discussed across various fields of the literature, from economics to geography, including models proposed by physicists. We propose here a general framework that allows to include different famous models (in particular Gibrat, Simon and Preferential Attachment model) within an unified vision. It brings first an insight into epistemological debates on the relevance of models. Furthermore, bridges between models lead to the possible transfer of analytical results to some models that are not directly tractable.

Seminal models of urban growth are Simon [SIMON, 1955] (later generalized as e.g. [HARAN et VINING, 1973]) and Gibrat models. **C : (Florent) à détailler davantage, c'est une matière basique de la thèse** Many examples can be given across disciplines. [BENGUIGUI et BLUMENFELD-LIEBERTHAL, 2007] give an equation-based dynamical model, whereas [GABAIX, 1999] solves a stationary model. **C : (Florent) après tu es dans l'implémentation A1 : non a priori, variantes et extensions** [GABAIX et IOANNIDES, 2004] reviews urban growth approaches in economics. A model adapted from evolutive urban theory is solved in [FAVARO et PUMAIN, 2011] and improves Gibrat models. The question of empirical scales at which it is consistent to study urban growth was also tackled in the particular case of France [BRETAGNOLLE, PAULUS et PUMAIN, 2002]. We stay to a certain level of tractability to include models as essence of our approach is links between models but do not make ontologic assumptions.

B.1.2 *Cadre de Travail*

what we propose as a framework can be understood as a meta-model in the sense of [COTTINEAU, CHAPRON et REUILLO, 2015], i.e. an modular general modeling process within each model can be understood as a limit case or as a specific case of another model. More simply it should be a diagram of formal relations between models. **C : (Florent) à ce stade on ne sait pas si tu vas faire 1 ou N modèles, c'est**

un choix qu'il te faut défendre avant d'en arriver là The ontological aspect is also tackled by embedding the diagram into an ontological state space (which discretization corresponds to the “bricks” of the incremental construction of [COTTINEAU, CHAPRON et REUILLOU, 2015]). It constructs a sort of model classification or modelography.

C : (Florent) PAS UTILE ICI JE PENSE

We are still at the stage of different derivations of links between models that are presented hereafter.

B.1.3 Dérivations

Généralisation de l'Attachement Préférentiel

[YAMASAKI et al., 2006] give a generalization of the classical Preferential Attachment Network Growth model, as a birth and death model with evolving entities. More precisely, network units gain and lose population (equivalent to links connexions) at fixed probabilities, and new unit can be created at a fixed rate.

Lien entre Gibrat et Attachement Préférentiel

C : (Florent) est-ce standard d'introduire de la stochasticité dans Gibrat : $P_{t+1} = R_P t$ A1 : c'est la formulation standard a priori Considérons un modèle de croissance strictement positive de Gibrat donnée par $P_i(t) = R_i(t) \cdot P_i(t-1)$ avec $R_i(t) > 1$, $\mu_i(t) = \mathbb{E}[R_i(t)]$ et $\sigma_i(t) = \mathbb{E}[R_i(t)^2]$. **C : (Florent) expliquer le sens des P,R etc.** D'autre part, soit un modèle simple d'attachement préférentiel, avec une probabilité d'attachement $\lambda \in [0, 1]$ et un nombre de nouveau arrivants $m > 0$. **C : (Florent) quelle est l'équation $P_{t+1} = P_t \cdot m \cdot \lambda$** Il est possible de dériver que le Gibrat est statistiquement équivalent à une limite de l'attachement préférentiel, sous l'hypothèse que toutes les fonctions génératrices des moments de $R_i(t)$ existent. Les distributions classiques qui peuvent être utilisées dans ce cas, e.g. une distribution normale ou log-normale, sont entièrement déterminées par leur deux premiers moments, ce qui rend cette hypothèse raisonnable. **C : (Florent) on a déjà discuté de cette eq Gibrat/att pref mais tu ne peux pas faire l'économie d'expliquer pourquoi tu t'es posé la question, i.e. à quoi cela va te servir ensuite**

Lemma 2 The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.

La preuve est donnée en Annexe ??.

C : (Florent) certain limit : à qualifier plus précisément

C : (Florent) je n'arrive pas à te suivre : si tu as besoin d'être relu sur ces développements, il faut convenir d'un rendez-vous pour que tu m'expliques le cheminement

Lien entre Simon et Attachement Préférentiel

A rewriting of Simon model yields a particular case of the generalized preferential attachment, in particular by vanishing death probability.

Lien entre Favarro-Pumain et Gibrat

[FAVARO et PUMAIN, 2011] generalizes Gibrat models with innovation propagation dynamics, being therefore a generalization of that model. Theoretically, a process-based model equivalent to the Favarro-Pumain should then fill the missing case in model classification at the corresponding discretization. Simpop models do not fill that case as they stay at the scale of city systems, as for Marius models [COTTINEAU, 2014]. These must also have their counterparts in discrete microscopic formulation.

C : (Florent) la encore tu parles de modèles que tu ne décris pas par ailleurs ; or Personnes connaissant FavaraoPumain \cap Personnes connaissant Gibrat $\cap \dots =$ quelques personnes sur terre !

Lien entre Bettencourt-West et Pumain

We are considering to study Bettencourt-West model for urban scaling laws [BETTENCOURT, LOBO et WEST, 2008] as entering the stochastic urban growth framework as stationary component of a random growth model, but investigation are still ongoing.

C : (Florent) on ne sait toujours pas dans quelle perspective tu fais cela

Autres modèles

[GABAIX, 1999] develops an economic model giving a Simon equivalent formulation. They in particular find out that in upper tail, proportional growth process occurs. We find the same result as a consequence of the derivation of the link between Gibrat and Preferential attachment models.

C : (Florent) je pense que tu as intérêt soit à présenter moins de modèles, mais plus en détails, soit à partir d'angles d'attaque précis et faire des typologies de modèles

B.2 SENSIBILITÉ DES LOIS D'ECHELLE URBAINES À L'ETENDUE SPATIALE

Au centre de la théorie évolutive des villes se trouvent la hiérarchie et les lois d'échelle associées. Nous proposons ici un bref développement méthodologique sur la sensibilité des lois d'échelle à la définition de la ville. **C : (Florent) présenté comme cela ce n'est pas évident de comprendre le rapport avec ta thèse**

Les lois d'échelle ont été montrées universelles des systèmes urbains à de nombreuses échelles et pour différents indicateurs. **C : (Florent) pas très précis** Des études récentes questionnent toutefois la cohérence de la détermination des exposants d'échelle, puisque leur valeur peut varier significativement selon les seuils utilisés pour définir les entités urbaines sur lesquelles les quantités urbaines sont intégrées, franchissant même dans certains cas la barrière qualitative de l'échelle linéaire, d'une loi infra-linéaire à une loi super-linéaire. Nous utilisons un modèle théorique simple de distribution spatiale des densités et des fonctions urbaines pour montrer analytiquement qu'un tel comportement peut être dérivé comme conséquence du type de distribution spatiale et de la méthode utilisée. Les simulations numériques confirment les résultats théoriques et révèle que les résultats sont raisonnablement indépendants du noyau spatial utilisé pour distribuer la densité.

Les lois d'échelle pour les systèmes urbains, en commençant par la bien connue loi rang-taille de Zipf pour la distribution des tailles des villes [GABAIX, 1999], **C : (Florent) déjà dit** ont été montrées être une caractéristique récurrente des systèmes urbains, à différentes échelles et pour différents types d'indicateurs. Elles reposent sur la constatation empirique que des indicateurs calculés sur des éléments du système urbain, qui peuvent être les villes dans le cas d'un système de villes, mais aussi des entités plus petites à une plus petite échelle, suivent relativement bien une distribution en loi de puissance en fonction de la taille de l'entité, i.e. pour l'entité i avec population P_i , on a pour une quantité intégrée A_i , la relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^\alpha$. Les exposants d'échelle α peuvent être plus petits ou plus grands que 1, menant à des effets infra ou supra-linéaires. Diverses interprétations thématiques de ce phénomène ont été proposées, typiquement sous la forme d'analyse des processus. La littérature économique contient une production abondante sur le sujet (voir [GABAIX et IOANNIDES, 2004] pour une revue), mais est généralement faiblement spatiale, donc de faible intérêt pour notre approche qui s'intéresse particulièrement à l'organisation spatiale. Des règles économiques simples comme un équilibre énergétique peut conduire à de simples lois d'échelles [BETTENCOURT, LOBO et WEST, 2008] mais sont difficiles à ajuster empiriquement. Une proposition intéressante par PUMAIN est qu'elles sont intrinsèquement dues au

caractère évolutionnaire des systèmes de villes, où l'émergence complexe par les interactions entre villes génère de telles distributions globales [PUMAIN et al., 2006]. Même si un parallèle tentant peut être fait avec les systèmes biologiques auto-organisés, PUMAIN insiste sur le fait que l'hypothèse d'ergodicité **C : (Florent) préciser ce que cela signifie** pour de tels systèmes n'est pas raisonnable dans le cas de systèmes géographiques et que l'analogie est difficilement exploitable [PUMAIN, 2012b]. D'autres explications ont été proposées à d'autres échelles, comme le modèle de croissance urbaine à échelle mesoscopique (échelle de la ville) donné dans [LOUF et BARTHELEMY, 2014] qui montre que la congestion dans les réseaux de transport pourrait être une raison de la forme des villes et des lois d'échelle correspondantes. On peut noter que les modèles "classiques" de croissance urbaine comme le modèle de Gibrat [FAVARO et PUMAIN, 2011] fournissent une approximation au premier ordre des systèmes exhibant des lois d'échelles, mais que les interactions entre agents doivent être incorporées dans le modèle pour obtenir un résultat plus fidèle aux données réelles, comme le modèle de Favaro-Pumain pour la propagation des cycles d'innovation proposé dans [FAVARO et PUMAIN, 2011], qui généralise un modèle de Gibrat pour la croissance des villes françaises avec une ontologie similaire à celle des modèles Simpop. **C : (Florent) ok : modèles qui reproduisent scaling, est-ce un des critères de validation des modèles que tu vas développer ?**

C : (Florent) qu'est ce que ça veut dire, blind application of models ?

The derivations in the simple case of exponential mixture density, are done in Appendix ??.

B.3 LIEN ENTRE CORRELATION SPATIO-TEMPORELLES STATIQUES ET DYNAMIQUES SOUS HYPOTHÈSES SIMPLIFIÉES

L'espace et le temps sont cruciaux pour l'étude des systèmes géographiques quand on cherche à comprendre les *processus* (par définition dynamiques [Hypergeo]) **C : (Florent) c'est déjà une lecture, certes processus renvoie à une évolution, mais les échelles de temps du modèle/processus ne sont pas nécessairement les mêmes** qui évoluent dans une structure spatiale au sens de [DOLLFUS et DASTÈS, 1975]. **C : (Florent) citer Cottineau ?**

[CROSS et HOHENBERG, 1994] : spatio-temporal chaos

The capture of neighborhood effects in statistical models is a wisely used practice in spatial statistics, as the technique of Geographically Weighted Regression illustrates [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998]. A possible interpretation among many definitions of spatial autocorrelation [GRIFFITH, 1992] yields that by estimating a plausible characteristic distance for spatial correlations or autocorrelations, one can isolate independent effects between variables from effects due to neighborhood interactions². The study of the spatial covariance structure is a cornerstone of advanced spatial statistics that was early formulated [GRIFFITH, 1980]. **C : (Florent) cela semble tout de même loin du sujet ou alors il faut que tu expliques clairement**
We propose now to study possible links between spatial and temporal correlations, using spatio-temporal covariance structure to infer information on dynamical processes.

B.3.1 Notations

We consider a multivariate spatio-temporal stochastic process denoted by $\vec{Y}[\vec{x}, t]$. At a given point \vec{x}_0 in space, we can define temporal covariance structure by

$$C_t(\vec{x}_0) = \text{Var}[\vec{Y}[\vec{x}_0, \cdot]]$$

and spatial covariance structure at fixed time by

$$C_x(t) = \text{Var}[\cdot, t]$$

It is clear that these quantities will be in practice first ill-defined because of the difficulty in interpreting such a process by a spatio-temporal random variable, secondly highly non-stationary in space and time. We stay however at a theoretical level to gain structural knowledge, **C : (Florent) sens ?** reviewing simple cases in which a formal link can be established.

² note that the formal link between models of spatial autocorrelation (see e.g. [GRIFFITH, 2012]) is not clear and should be further investigated

B.3.2 Equation des Ondes

C : (Florent) pourquoi aborder cela ? A1 : cas idéal des STARMA, ondes d'innovation etc. : approche fondamentalement liée à l'analyse spatiale, mais bien plus complexe qu'une simple équation. justifie cette approche de lien spatio-temporel ?

In the case of propagating waves, there is an immediate link. Let assume that a wave equation if verified by "deterministic" parts of components

$$c^2 \cdot \partial_t^2 \bar{Y}_i = \Delta \bar{Y}_i \quad (13)$$

with $Y_i = \bar{Y}_i + \varepsilon_i$. If errors are uncorrelated and processes are stationary, we have then directly

$$\mathbf{C}_t [\partial_t^2 Y_i, \partial_t^2 Y_j] = \frac{1}{c^2} \cdot \mathbf{C}_x [\Delta Y_i, \Delta Y_j] \quad (14)$$

This gives us however few insight on real systems as local diffusion, stationary assumptions and uncorrelated noises are far from being verified in empirical situations.

B.3.3 Equation de Fokker-Planck

An other interesting approach may when the process verifies a Fokker-Planck equation on probabilities of the state of the system when it is given by its position (diffusion of particles in that case)

$$\partial_t P(x_i, t) = -d \cdot \partial_x P(x_i, t) + \frac{\sigma^2}{2} \partial_x^2 P(x_i, t) \quad (15)$$

With no cross-correlation terms in the Fokker-Planck equation, covariance between processes vanish. We have finally in that case only a relation between averaged spatial and temporal variances that brings no information to our question.

B.3.4 Equation Maitresse

In the case of a master equation on probabilities of discrete states of the system

$$\partial_t \vec{P} = W \vec{P} \quad (16)$$

we have then for state i , $\partial_t P_i = \sum_j W_{ij} P_j$. As this relation is at a fixed time we can average in time to obtain an equation on temporal covariance. It is not clear how to make the link with spatial covariance as these will depend on spatial specification of discrete states. This question is still under investigation.

B.3.5 Echantillonnage spatial cohérent

In a more empirical way, we propose to not assume any constraint of process dynamics but to however investigate how the computation of spatial correlations can inform on temporal correlations. We try to formulate easily verifiable assumptions under which this is possible.

We make the following assumptions on the spatio-temporal stochastic processes $Y_i[\vec{x}, t]$:

1. Local spatial autocorrelation is present and bounded by l_ρ (in other words the processes are continuous in space) : at any \vec{x} and t , $|\rho_{\|\Delta\vec{x}\| < l_\rho} [Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$. **C : (Florent) je ne comprends pas ce qui est écrit, qu'une abs soit > 0 ok, donc c'est autre mais quoi? A1 : c'est le strict > 0 qui compte, c'est une façon de postuler que les processus sont continus à une certaine échelle fine**
2. Processes are locally parametrized : $Y_i = Y_i[\alpha_i]$, where $\alpha_i(\vec{x})$ varies with l_α , with $l_\alpha \gg l_\rho$.
3. Spatial correlations between processes have a sense at an intermediate scale l such that $l_\alpha \gg l \gg l_\rho$.
4. Processes covariance stationarity times scale as \sqrt{l} .
5. Local ergodicity is present at scale l and dynamics are locally chaotic.

Assumptions one to three can be tested empirically and allow to compare spatial correlation estimated on spatial samplings at scale l . Assumption four is more delicate as we are precisely constructing this methodology because we have no temporal information on processes. It is however typical of spatial diffusion processes, and population or innovation diffusion should verify this assumption. **C : (Florent) cela devrait être un point de départ (expliquerait pourquoi ces modèles ; te ferait peut être en considérer d'autres** The last assumption can be tested if feasible space is known, by checking cribbing on image space on the spatial sample. Under these conditions, local spatial sampling is equivalent to temporal sampling and spatial correlation estimators provide estimator of temporal correlations.

B.4 GÉNÉRATION DE DONNÉES SYNTHÉTIQUES CORRÉLÉES

La génération de données synthétiques hybrides similaires à des données réelles présente des enjeux méthodologiques et thématiques pour la plupart des disciplines dont l'objet est l'étude de systèmes complexes. Comme l'interdépendance entre les éléments constitutifs d'un système, matérialisée par leur relations, conduit à l'émergence de ses propriétés macroscopiques, une possibilité de contrôle de l'intensité des dépendances dans un jeu de données synthétiques est un instrument de connaissance du comportement du système. Nous proposons une méthodologie de génération de données synthétiques hybrides sur lequel la structure de correlation est contrôlée. La méthode est illustrée sur des séries temporelles financières et permet l'étude de l'interférence entre composantes à différentes fréquences sur la performance d'un modèle prédictif, en fonction des correlations entre composantes à différentes échelles. On présente ensuite une application à un système géographique, dans laquelle le couplage faible d'un modèle de distribution de densité de population avec un modèle de génération de réseau permet la simulation de configurations territoriales, qui sont calibrées selon des objectifs morphologiques sur l'ensemble de l'Europe. L'exploration intensive du modèle permet l'obtention d'un large spectre de valeurs pour la matrice de correlation entre mesures morphologiques et mesures du réseau. On démontre ainsi les possibilités d'applications variées et les potentialités de la méthode.

B.4.1 Contexte

L'utilisation de données synthétiques, au sens de populations statistiques d'individus générées aléatoirement sous la contrainte de reproduire certaines caractéristiques du système étudié, est une pratique méthodologique largement répandue dans de nombreuses disciplines, et particulièrement pour des problématiques liées aux systèmes complexes, telles que par exemple l'évaluation thérapeutique [ABADIE, DIAMOND et HAINMUELLER, 2010], l'étude des systèmes territoriaux [MOECKEL, SPIEKERMANN et WEGENER, 2003; PRITCHARD et MILLER, 2009], l'apprentissage statistique [BOLÓN-CANEDO, SÁNCHEZ-MAROÑO et ALONSO-BETANZOS, 2013] ou la bio-informatique [BULCKE et al., 2006]. Il peut s'agir d'une désagrégation par création d'une population au niveau microscopique présentant des caractéristiques macroscopiques données, ou bien de la création de nouvelles populations au même niveau d'agrégation qu'un échantillon donné avec un critère de ressemblance aux données réelles. Le niveau de ce critère peut dépendre des applications attendues et peut par exemple aller de la fidélité des distributions statistiques pour un certain nombre d'indicateurs à des contraintes plus faibles de valeurs pour des indica-

teurs agrégés, c'est à dire l'existence de motifs macroscopiques similaires. Dans le cas de systèmes chaotiques ou présentant de fortes caractéristiques d'émergence, une contrainte microscopique n'implique pas nécessairement le respect des motifs macroscopiques, et arriver à les reproduire est justement un des enjeux des pratiques de modélisation et simulation en sciences de la complexité. La donnée, qu'elle soit simulée, mesurée ou hybride est au cœur de l'étude des systèmes complexes de par la maturation de nouvelles approches computационnelles [ARTHUR, 2015], il est donc essentiel d'étudier des procédures d'extraction d'information des données (fouille de données) et de simulation d'une information similaire (génération de données synthétiques).

Si le premier ordre est de manière générale bien maîtrisé, il n'est pas systématique ni aisément de contrôler le second ordre, c'est à dire les structures de covariance entre les variables générées, même si des exemples spécifiques existent, comme dans [YE, 2011] où la sensibilité des sorties de modèles de choix discrets à la forme des distributions des variables aléatoires ainsi qu'à leur structures de dépendance. Il est également possible d'interpréter les modèles de génération de réseaux complexes [NEWMAN, 2003] comme la création d'une structure d'interdépendance au sein d'un système, représentée par la topologie des liens. Nous proposons ici une méthode générique prenant en compte l'interdépendance lors de la génération de données synthétiques, sous la forme de correlations.

L'ensemble des méthodologies mentionnées en introduction sont trop variées pour être résumées par un même formalisme. Nous proposons ici une formulation générique ne dépendant pas du domaine d'application, ciblée sur le contrôle de la structure de correlation des données synthétiques.

B.4.2 Formalisation

Soit un processus stochastique multidimensionnel \vec{X}_I (l'ensemble d'indexation pouvant être par exemple le temps dans le cas de séries temporelles, l'espace, ou une indexation quelconque). On se propose, à partir d'un jeu de réalisations $\mathbf{X} = (X_{i,j})$, de générer une population statistique $\tilde{\mathbf{X}} = \tilde{X}_{i,j}$ telle que

1. d'une part un certain critère de proximité aux données est vérifié, i.e. étant donné une précision ε et un indicateur f , $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
2. d'autre part le niveau de correlation est contrôlé, i.e. étant donné une matrice fixant une structure de covariance R , $\text{Var}[(\tilde{X}_i)] = R$, où la matrice de variance/covariance est estimée sur la population synthétique.

La satisfaction du deuxième point sera généralement conditionnée par la valeur de paramètres, dont dépendra la procédure de génération, qu'il s'agisse de modèles simples ou complexes. Formellement, les processus synthétiques sont des familles paramétriques $\tilde{X}_i[\vec{\alpha}]$. Nous proposons de décliner cette méthode sur deux exemples très différents mais tous deux typiques des systèmes complexes : des séries temporelles financières à haute fréquence, et les systèmes territoriaux. On illustre ainsi la flexibilité de la logique, ouvrant des portes interdisciplinaires par l'exportation de méthodes ou raisonnements par exemple. Dans le premier cas, la proximité aux données est l'égalité des signaux à une fréquence fondamentale, auxquels on superpose des composantes synthétiques dont il est facile de contrôler le niveau de correlation. On se place dans une logique de données hybrides, pour tester des hypothèses ou modèles dans un contexte plus proche de la réalité que sur des données purement synthétiques. Cet exemple, sans rapport thématique avec la thèse, est présenté en Appendice C.2. Dans le deuxième cas, la calibration morphologique d'un modèle de distribution de densité de peuplement permet de respecter le critère de proximité aux données. Les correlations de la forme urbaine avec celle d'un réseau de transport sont ensuite obtenues empiriquement par exploration du couplage avec un modèle de génération de réseau. Leur contrôle est dans ce cas indirect puisque constaté empiriquement.

UNE VUE ALTERNATIVE : DONNÉES SYNTHÉTIQUES Let M_m a stochastic model of simulation, which inputs are to simplify initial conditions D_0 and parameters $\vec{\alpha}$, and output $M_m[\vec{\alpha}, D_0](t)$ at a given time t . We assume that it is partially data-driven in the sense that D_0 is supposed to represent a real situation at a given time, and model performance is measured by the distance of its output at final time to the real situation at the corresponding time, i.e. error function is of the form $\|\mathbb{E}[\vec{g}(M_m[\vec{\alpha}, D_0](t_f))] - \vec{g}(D_f)\|$ where \vec{g} is a deterministic field corresponding to given indicators.

Evaluating the model on real data is rapidly limited in control possibilities, being restricted to the search of datasets allowing natural control groups. Furthermore, statistical behaviors are generally poorly characterized because of the small number of realizations. Working with synthetic data first allows to solve this issue of robustness of statistics, and then gives possibilities of control on some “meta-parameters” in the sense described before.

B.5 UN CADRE BASÉ SUR LA DISCRÉPANCE POUR COMPARER LA ROBUSTESSE DES EVALUATIONS MULTI-ATTRIBUTS

Les évaluations multi-objectifs sont un aspect essentiel de la gestion de systèmes complexes, puisque la complexité intrinsèque d'un système est généralement étroitement liée au nombre d'objectifs d'optimisation potentiels. Cependant, une évaluation ne fait pas sens si sa robustesse, au sens de sa fiabilité, n'est pas donnée. Les méthodes statistiques usuelles fournissant une mesure de robustesse sont très dépendantes des modèles sous-jacents. Nous proposons une formulation d'un cadre indépendant du modèle, dans le cas d'indicateurs intégrés et agrégés (évaluation multi-attributs), qui permet de définir une mesure de robustesse relative prenant en compte la structure des données et les valeurs des indicateurs. La méthode est testée sur données urbaines synthétiques associées aux arrondissements de Paris, et à des données réelles de revenus pour l'évaluation de la ségrégation urbaine dans la région métropolitaine du Grand Paris. Les premiers résultats numériques montrent les potentialités de cette nouvelle méthode. De plus, sa relative indépendance au type de système et au modèle pourrait la positionner comme une alternative aux méthodes statistiques classiques d'évaluation de la robustesse.

B.5.1 *Introduction*

Contexte Général

Les problèmes multi-objectifs sont organiquement liés à la complexité des systèmes sous-jacents. En effet, que ce soit dans le champ des *Systèmes Complexes Industriels*, dans le sens de systèmes conçus par ingénierie, où la construction de Systèmes de Systèmes (SoS) par couplage et intégration induit souvent des objectifs contradictoires [MARLER et ARORA, 2004], ou dans le champ des *Systèmes Complexes Naturels*, au sens de systèmes non désignés, physiques, biologiques ou sociaux, qui présentent des propriétés d'émergence et d'auto-organisation, pour lesquels les objectifs peuvent e.g. être le résultat de l'interaction d'agents hétérogènes (voir [NEWMAN, 2011] pour une revue étendue des types de systèmes concernés par cette approche), l'optimisation multi-objectifs peut être explicitement introduite pour étudier ou désigner le système, mais régit généralement déjà implicitement les mécanismes internes du système. Le cas des Systèmes Complexes Sociaux-techniques est particulièrement intéressant puisque selon Haken [HAKEN et PORTUGALI, 2003], ils peuvent être vus comme des systèmes hybrides embarquant des agents sociaux dans des "artefacts techniques" (parfois jusqu'à un niveau inattendu, créant ce que PICON décrit comme *cyborts* [PICON, 2013]), et cumulent ainsi la potentialité d'être à l'origine

de problèmes multi-objectifs³. La notion récente d'*éco-quartier* [SOUAMI, 2012] est un exemple typique pour lequel la durabilité implique des objectifs contradictoires. L'exemple des systèmes de transport, dont la conception a glissé durant la seconde moitié du 20ème siècle d'analyses coût-bénéfices à la price de décision multi-critères, est également typique de tels systèmes [BAVOUX et al., 2005]. Les systèmes géographiques sont à présent bien étudiés d'un tel point de vue, en particulier grâce à l'intégration des cadres multi-objectifs au sein des Systèmes d'Information Géographiques [CARVER, 1991]. Comme dans le cas microscopique des éco-quartiers, la planification et le design urbains mésoscopiques et macroscopiques peuvent être rendus durables grâce aux évaluations par indicateurs [JÉGOU et al., 2012].

Un aspect crucial de l'évaluation est une certaine notion de sa fiabilité, que nous nommerons ici *robustesse*. Les méthodes statistiques incluent naturellement cette notion puisque la construction et l'estimation de modèles statistiques donne divers indicateurs de la consistance des résultats [LAUNER et WILKINSON, 2014]. Le premier exemple venant à l'esprit est l'application de la loi des grands nombres pour obtenir la *p-valeur* d'une estimation de modèle, qui peut être interprété comme une mesure de confiance en les valeurs estimées. D'autre part, les intervalles de confiance et le *beta-power* sont d'autres indicateurs importants de robustesse statistique. L'inférence bayésienne fournit également des mesures de robustesse quand la distribution des paramètres est estimée de manière séquentielle. Concernant les optimisations multi-objectifs, en particulier par des algorithmes heuristiques (comme par exemple les algorithmes génétiques, ou les solveurs de recherche opérationnelle), la notion de robustesse d'une solution consiste plus en la stabilité de la solution dans l'espace des phases du système dynamique correspondant. Des progrès récents ont été faits vers une formulation unifiée de la robustesse pour les problèmes d'optimisation multi-objectifs, comme dans [DEB et GUPTA, 2006] où les fronts de Pareto robustes sont définis comme des solutions insensibles aux petites perturbations. Dans [BARRICO et ANTUNES, 2006], la notion de degré de robustesse est introduite, formalisée comme une sorte de continuité des autres solutions dans des voisinages successifs d'une solution.

Cependant, il n'existe pas de méthode générique qui permettrait une évaluation de la robustesse de façon indépendante au modèle, i.e. qui serait extraite de la structure des données et des indicateurs mais ne dépendrait pas de la méthode utilisée. Un avantage serait par exemple une estimation *a priori* de la robustesse potentielle d'une évaluation et de décider ainsi si elle vaut la peine d'être faite. Nous proposons un cadre répondant à cette contrainte dans le cas particu-

³ Nous désignons ici par *Evaluation Multi-objectifs* toutes les pratiques incluant le calcul de multiples indicateurs d'un système (il peut s'agir d'optimisation multi-objectif pour un design de système, une évaluation multi-objectif d'un système existant, une évaluation multi-attributs ; notre cadre particulier correspondra au dernier cas).

lier des évaluations multi-attributs, i.e. quand le problème est rendu unidimensionnel par agrégation des objectifs. Il est basé sur les données et non sur les modèles, au sens où l'estimation de la robustesse ne dépendra pas de la manière dont les indicateurs sont calculés, tant qu'ils respectent certaines hypothèses détaillées par la suite.

Approche Proposée

OBJECTIFS COMME INTÉGRALES SPATIALES Nous supposons que les objectifs peuvent être exprimés comme intégrales spatiales, ce qui devrait s'appliquer à tout système territorial, et nos cas d'application sont des systèmes urbains. Ce n'est pas si restrictif en terme d'indicateurs possibles si l'on utilise les bonnes variables et noyaux intégrés : de façon analogue à la méthode de Regression Géographique Pondérée [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998], toute variable spatiale peut être intégrée contre des noyaux réguliers de taille variable et le résultats sera une agrégation spatiale dont la signification dépendra de l'étendue du noyau. Les exemples utilisés par la suite comme des moyennes conditionnelles ou des sommes vérifient parfaitement cette hypothèse. Même un indicateur déjà agrégé dans l'espace peut être interprété comme une intégrale spatiale en utilisant une distribution de Dirac au centroïde de la zone correspondante.

OBJECTIFS AGRÉGÉS LINÉAIREMENT Une seconde hypothèse que nous faisons est que l'évaluation multi-objectifs est effectuée par agrégation linéaire des objectifs, c'est à dire qu'on se place dans le cadre d'un problème d'optimisation multi-attributs. Si $(q_i(\vec{x}))_i$ sont les valeurs des fonctions objectifs, on définit alors des poids $(w_i)_i$ afin de construire la fonction de prise de décision $q(\vec{x}) = \sum_i w_i q_i(\vec{x})$, dont la valeur détermine ensuite la performance d'une solution. Cette approche est analogue aux utilités agrégées en économie et est utilisée dans de nombreux domaines. La subtilité réside dans le choix des poids, i.e. de la forme de la fonction de projection, et différentes solutions ont été développées pour obtenir des poids selon la nature du problème. Récemment, [DOBBIE et DAIL, 2013] a proposé de comparer la robustesse des différentes techniques d'agrégation par une analyse de sensibilité, effectuée par simulations de Monte-Carlo pour produire des données synthétiques, ce qui permet d'obtenir la distribution des biais pour les différentes techniques, certaines étant significativement plus performantes que d'autres. Toutefois, la quantification de la robustesse dépend toujours des modèles utilisés dans ce travail.

Le reste de cette monographie est organisé de la façon suivante : la section 2 décrit intuitivement puis mathématiquement le cadre proposé ; la section 3 détaille ensuite l'implémentation, la collecte des données pour les cas d'étude et les résultats numériques pour une évaluation intra-urbaine synthétique et un cas réel métropolitain ; la

section 4 discute finalement les limitations et les potentialités de la méthode.

B.5.2 *Description du Cadre*

Description Intuitive

Nous décrivons à présent le cadre proposé pour permettre théoriquement de comparer la robustesse d'évaluation de deux systèmes urbains différents. Ce cadre est une généralisation d'une méthode empirique proposée dans [ALI et al., June 2014] pour accompagner une étude dans un autre contexte effectuant une comparaison du sens et de la pertinence des indicateurs dans un contexte de durabilité. Intuitivement, la base empirique se base sur les principes suivants :

- Les systèmes urbains peuvent être vus selon l'information disponible, i.e. les données brutes décrivant le système. Dans une approche basée sur les données, celles-ci sont la base de notre cadre et la robustesse sera déterminée par leur structure.
- A partir des données sont capturés des indicateurs (fonctions objectifs). Nous supposons qu'un choix d'indicateurs est une intention particulière de traduire des aspects particuliers du système, i.e. de capturer une réalisation d'un "fait urbain" au sens de MANGIN [MANGIN et PANERAI, 1999] - une sorte de fait stylisé en terme de processus et de mécanismes, ayant différentes réalisations sur des systèmes distincts dans l'espace, dépendant de chaque contexte géographique précis.
- Etant donné plusieurs systèmes et indicateurs associés, un espace commun peut être construit pour les comparer. Dans cet espace, les données représentent plus ou moins bien le système réel, c'est à dire qu'elles sont imprécises en fonction de l'échelle initiale, de la précision effective des données. Nous proposons de capturer exactement ces différents aspects au travers de la notion de discrépance d'un nuage de points, qui est un outil mathématique provenant des théories d'échantillonnage, permettant d'exprimer la façon dont un jeu de données rempli l'espace dans lequel il s'insère [DICK et PILLICHSHAMMER, 2010].

Synthétisant ces contraintes, nous proposons une notion de *Robustesse* d'une évaluation qui capture à la fois, en combinant la fiabilité des données à l'importance relative des indicateurs,

1. *Données manquantes* : une évaluation se basant sur des jeux de données plus raffinés sera naturellement plus robuste.
2. *Importance des indicateurs* : les indicateurs avec plus d'importance relative pèseront plus dans la robustesse totale.

Description Formelle

INDICATEURS Soit $(S_i)_{1 \leq i \leq N}$ un nombre fini de systèmes territoriaux géographiquement disjoints, que nous supposons décrits par les données brutes et des indicateurs intermédiaires, donnés par $S_i = (\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{X}_i \times \mathcal{Y}_i$ avec $\mathcal{X}_i = \prod_k \mathcal{X}_{i,k}$ tel que chaque sous-espace contient des matrices réelles : $\mathcal{X}_{i,k} = \mathbb{R}^{n_{i,k}^X p_{i,k}^X}$ (de la même façon pour \mathcal{Y}_i). Nous définissons également une fonction d'indice ontologique $I_X(i, k)$ (resp. $I_Y(i, k)$) prenant des valeurs entières qui coincident si et seulement si les deux variables ont même ontologie au sens de [LIVET et al., 2010], c'est à dire qu'elles sont supposées représenter le même objet réel. On distingue les "données brutes" \mathbf{X}_i à partir desquelles les indicateurs sont calculés généralement par des fonctions déterministes explicites, des "indicateurs intermédiaires" \mathbf{Y}_i qui sont déjà intégrés et peuvent être par exemple les sorties de modèles élaborés simulant certains aspects du système urbain. Nous définissons l'espace caractéristique du "fait urbain" par

$$(\mathcal{X}, \mathcal{Y}) \underset{\text{def}}{=} \left(\prod \tilde{\mathcal{X}}_c \right) \times \left(\prod \tilde{\mathcal{Y}}_c \right) = \left(\prod_{\mathcal{X}_{i,k} \in \mathcal{D}_X} \mathbb{R}^{p_{i,k}^X} \right) \times \left(\prod_{\mathcal{Y}_{i,k} \in \mathcal{D}_Y} \mathbb{R}^{p_{i,k}^Y} \right) \quad (17)$$

avec $\mathcal{D}_X = \{\mathcal{X}_{i,k} | I(i, k) \text{ distincts, } n_{i,k}^X \text{ maximal}\}$ (de même pour \mathcal{Y}_i). Il s'agit en fait de l'espace abstrait sur lequel les indicateurs sont intégrés. Les indices c introduit par définition correspondent aux différents indicateurs au sein des différents systèmes. Cette espace est l'espace minimal commun à tous les systèmes permettant une définition commune des indicateurs pour tous.

Soit $\mathbf{X}_{i,c}$ les données projetées canoniquement sur le sous-espace correspondant, bien définies pour tout i et tout c . Nous faisons donc l'hypothèse clé que tous les indicateurs sont calculés par intégration contre un noyau donné, i.e. pour tout c il existe H_c espace de fonctions à valeurs réelles sur $(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)$, tel que pour tout $h \in H_c$:

1. h est "suffisamment" régulière (distribution tempérée par exemple)
2. $q_c = \int_{(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)} h$ est une fonction décrivant le "fait urbain" (l'indicateur en lui-même)

Des exemples typiques de noyaux peuvent être :

- Une moyenne des lignes de $\mathbf{X}_{i,c}$ est calculée par $h(x) = x \cdot f_{i,c}(x)$ où $f_{i,c}$ est la densité de la distribution de la variable sous-jacente.
- Un taux d'éléments du jeu de données respectant une condition donnée C , $h(x) = f_{i,c}(x) \chi_{C(x)}$.

- Pour des variables déjà agrégées \mathbf{Y} , une distribution de Dirac permet de les exprimer également comme des intégrales de noyaux.

AGRÉGATION La détermination des poids est en fait le point crucial des processus de prise de décision multi-attributs, et de nombreuses méthodes sont disponibles (voir [WANG et al., 2009] pour une revue dans le cas particulier de la gestion de l'énergie durable). Définissons les poids pour l'agrégation linéaire. Nous supposons les indicateurs normalisés, i.e. $q_c \in [0, 1]$, pour une construction plus simple des poids relatifs. Pour i, c et $h_c \in H_c$ donnés, le poids $w_{i,c}$ est simplement constitué par l'importance relative de l'indicateur $w_{i,c}^L = \frac{\hat{q}_{i,c}}{\sum_c \hat{q}_{i,c}}$ où $\hat{q}_{i,c}$ est un estimateur de q_c pour les données $X_{i,c}$ (i.e. la valeur calculée effectivement). On peut noter que cette étape n'est pas contrainte et que cela peut être étendu à tout ensemble d'attribution de poids, en prenant par exemple $\tilde{w}_{i,c} = w_{i,c} \cdot w'_{i,c}$ si \mathbf{w}' sont les poids fixés par le preneur de décisions. Nous nous concentrerons sur l'influence relative des attributs et pour cela choisissons cette forme simple pour les poids.

ESTIMATION DE LA ROBUSTESSE La scène est à présent apprêtée pour construire une estimation de la robustesse d'une évaluation faite par la fonction d'agrégation. Pour cela, nous appliquons un théorème d'approximation d'intégrale similaire au méthodes introduites dans [VARET, 2010], puisque la forme intégrée des indicateurs permet justement de bénéficier de tels résultats théoriquement puissant. Soit $\mathbf{X}_{i,c} = (\vec{X}_{i,c,l})_{1 \leq l \leq n_{i,c}}$ et $D_{i,c} = \text{Disc}_{\tilde{X}_c, L^2}(\mathbf{X}_{i,c})$ le discrépance du jeu de données⁴ [NIEDERREITER, 1972]. Avec $h \in H_c$, on a la borne supérieure sur l'erreur d'approximation de l'intégrale

$$\left\| \int h_c - \frac{1}{n_{i,c}} \sum_l h_c(\vec{X}_{i,c,l}) \right\| \leq K \cdot \|h_c\| \cdot D_{i,c}$$

où K est une constante indépendante des points de données et des fonctions objectifs. Cela donne directement

$$\left\| \int \sum w_{i,c} h_c - \frac{1}{n_{i,c}} \sum_l w_{i,c} h_c(\vec{X}_{i,c,l}) \right\| \leq K \sum_c |w_{i,c}| \|h_c\| \cdot D_{i,c}$$

En supposant l'erreur réalisée de manière raisonnable (scénario du "pire de cas" pour la connaissance de la valeur théorique de la fonc-

⁴ La discrépance est définie comme la norme-L2 de la discrépance locale qui est pour des points de données normalisés $\mathbf{X} = (x_{ij}) \in [0, 1]^d$, une fonction de $t \in [0, 1]^d$ comparant le nombre de points compris dans le volume de l'hypercube correspondant, donné par $\text{disc}(t) = \frac{1}{n} \sum_i \mathbb{1}_{\prod_j x_{ij} < t_j} - \prod_j t_j$. C'est une mesure de la manière dont le nuage de points couvre l'espace.

tion agrégée), nous prenons cette borne supérieure comme une approximation de sa magnitude. De plus, la normalisation des indicateurs implique que $\|\mathbf{h}_c\| = 1$. Nous proposons alors de comparer les bornes d'erreurs entre deux évaluations. Elle dépendent seulement de la distribution des données (équivalence à la *robustesse statistique*) et des indicateurs choisis (sorte de *robustesse ontologique*, i.e. est-ce que les indicateurs ont un sens réel dans le contexte choisi et est-ce que leur valeur fait sens), et sont un moyen de combiner ces deux types de robustesse dans une seule valeur.

Nous définissons ainsi un *ratio de robustesse* pour comparer la robustesse de deux évaluations par

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \quad (18)$$

L'interprétation intuitive de cette définition est que l'on compare la robustesse des évaluations en comparant la plus grande erreur faite dans chaque cas selon la structure des données et l'importance relative.

En construisant une relation d'ordre sur les évaluations en comparant la position du ratio par rapport à un, il est clair qu'on obtient un ordre complet sur l'ensemble des évaluations possibles. Ce ratio devrait en théorie permettre de comparer n'importe quelle évaluation d'un système urbain. Afin de garder un sens ontologique à cela, il devrait être utilisé pour comparer des sous-systèmes disjoints avec une proportion raisonnable d'indicateurs en commun, ou le même sous-système avec des indicateurs différents. On peut noter que cela fournit un moyen de tester l'influence des indicateurs sur une évaluation, en analysant la sensibilité du ratio à leur suppression. Au contraire, la détermination d'un nombre "minimal" d'indicateurs faisant chacun varier le ratio fortement pourrait être un moyen d'isoler des paramètres essentiels régissant le sous-système.

B.5.3 Résultats

IMPLÉMENTATION Le pré-traitement des données géographiques est fait via QGIS [QGIS, 2011] pour des raisons de performances. L'implémentation du coeur est faite en R [TEAM, 2000] pour la flexibilité de la gestion des données et du traitement statistique. De plus, le package DiceDesign [FRANCO et al., 2009] conçu pour les expériences numériques et l'échantillonnage, permet un calcul efficient et direct des discordances. Enfin, tout aussi important, l'ensemble du code source est disponible de manière ouverte sur le dépôt git du projet⁵ pour permettre la reproductibilité et la réutilisation [RAM, 2013].

⁵ à <https://github.com/JusteRaimbault/RobustnessDiscrepancy>

Implémentation sur Données Synthétiques

Nous proposons dans un premier temps d'illustrer l'implémentation par une application à des données et indicateurs synthétiques, pour des indicateurs de qualité de vie intra-urbaine pour la ville de Paris.

COLLECTE DES DONNÉES Le cas virtuel se base sur des données géographiques réelle, en particulier pour les arrondissements parisiens. Nous utilisons les données disponibles par le projet OpenStreetMap [BENNETT, 2010] qui fournit déjà des données précises en haute définition pour de nombreux aspects urbains. Nous utilisons le réseau de rues et la position des bâtiments dans la ville de Paris. Les limites des arrondissements, utilisées pour agréger et extraire les features lorsqu'on travaille sur un seul district, sont aussi pris de la même source. Nous utilisons les centroïdes des polygones des bâtiments et les segments du réseau de rues. Le jeu de données brutes consiste d'environ 200k bâtiments et 100k segments de rues.

CAS VIRTUEL Nous travaillons sur chaque arrondissement de Paris (du 1^{er} au 20^{ème}) comme un système urbain évalué. Des données synthétiques aléatoires sont associées aux features spatiales, chaque arrondissement pouvant alors être évalué de manière stochastique, et des répétitions permettent d'obtenir le comportement statistique moyen des indicateurs jouets et des ratios de robustesse. Les indicateurs choisis doivent être calculés comme des indicateurs résidentiels et du réseau de rues. Pour montrer différents exemples, nous implementons deux kernels moyens et une moyenne conditionnelle, tous liés à la durabilité environnementale et la qualité de vie, chacun devant être maximisés. On peut noter que ces indicateurs ont un sens réel mais pas de raison particulière d'être agrégés, ils sont ici choisis pour l'aspect pratique du modèle jouet et de la génération de données synthétiques. Avec $a \in \{1 \dots 20\}$ le nombre d'arrondissements, $A(a)$ l'aire spatiale correspondante à chacun, $b \in B$ les coordonnées des bâtiments et $s \in S$ les segments de rues, nous prenons

- Le complémentaire de la distance journalière moyenne au travail en voiture par individu, approché par, avec $n_{cars}(b)$ nombre de voiture dans le bâtiment (généré aléatoirement en associant des voitures à bâtiments proportionnel au taux de motorisation attendu α_m 0.4 à Paris), d_w distance des individus à leur travail (généré à partir du bâtiment vers un point aléatoire distribué uniformément dans l'étendue spatiale du jeu de données), et d_{max} le diamètre de l'aire de Paris, $\bar{d}_w = 1 - \frac{1}{|b \in A(a)|} \cdot \sum_{b \in A(a)} n_{cars}(b) \cdot \frac{d_w}{d_{max}}$
- Le complémentaire des flots moyens de voitures des rues dans la zone, approché par, avec $\varphi(s)$ flot relatif dans le segment de rue s , généré par le minimum entre 1 et une distribution

log-normale ajustée pour avoir 95% de masse plus petite que 1, ce qui mimique la distribution hiérarchique de l'utilisation des rues (qui correspond à la centralité de chemin), et $l(s)$ longueur du segment, $\bar{\varphi} = 1 - \frac{1}{|A(a)|} \cdot \sum_{s \in A(a)} \varphi(s) \cdot \frac{l(s)}{\max(l(s))}$

- Longueur relative de rues piétonnes \bar{p} , calculé via une dummy variable aléatoire uniforme ajustée pour obtenir une proportion fixée de segments pédestre.

Comme les données synthétiques sont stochastiques, les simulations sont lancées pour chaque quartier $N = 50$ fois, ce qui était un compromis raisonnable entre convergence statistique et temps nécessaire au calcul. La table 1 montre les résultats (moyennes et déviations standard) des valeurs des indicateurs et le calcul du ratio de robustesse. Les déviations standard obtenues confirment que ce nombre de simulations donnent des résultats consistants. Les indicateurs obtenus en fixant un ratio fixe montrent peu de variabilité, ce qui peut être une limite de cette approche jouet. On obtient toutefois le résultat intéressant que la majorité des arrondissements donne des évaluations plus robustes que le 1er arrondissement, ce qui était attendu par la taille et la fonction de ce quartier : il s'agit en effet d'un petit quartier avec de grand bâtiment administratifs, ce qui implique moins d'éléments spatiaux et pour cela une évaluation moins robuste selon la définition qu'on en a donnée.

Application à un cas réel : ségrégation métropolitaine

Le premier exemple avait pour but de montrer les potentialités de la méthode mais était purement synthétique, ne pouvant pour cela fournir pas de conclusion concrète ni d'implications pour la gouvernance. Nous proposons maintenant de l'appliquer à des données réelles dans le cas de la ségrégation métropolitaine.

DONNÉES Nous travaillons sur les données de revenus, disponible pour la France à un niveau intra-urbain (unités statistiques élémentaires IRIS) pour l'année 2011 sous la forme de résumé statistiques (déciles uniquement si la zone est peuplée suffisamment pour assurer l'anonymat), fournies par l'INSEE⁶. Les données sont associées à l'étendue géographique des unités statistiques, permettant le calcul d'indicateurs d'analyse spatiale.

INDICATEURS Nous utilisons ici trois indicateurs de ségrégation intégrés sur une zone géographique. Supposons la zone divisée en unités couvrantes S_i pour $1 \leq i \leq N$ avec pour centroïdes (x_i, y_i) . Chaque unité a des caractéristiques de population P_i et de revenu

⁶ <http://www.insee.fr>

Arrdt	$\langle \bar{d}_w \rangle \pm \sigma(\bar{d}_w)$	$\langle \bar{\varphi} \rangle \pm \sigma(\bar{\varphi})$	$\langle \bar{p} \rangle \pm \sigma(\bar{p})$	$R_{i,1}$
1 th	0.731655 ± 0.041099	0.917462 ± 0.026637	0.191615 ± 0.052142	1.000000 ± 0.0000
2 th	0.723225 ± 0.032539	0.844350 ± 0.036085	0.209467 ± 0.058675	1.002098 ± 0.0399
3 th	0.713716 ± 0.044789	0.797313 ± 0.057480	0.185541 ± 0.065089	0.999341 ± 0.0488
4 th	0.712394 ± 0.042897	0.861635 ± 0.030859	0.201236 ± 0.044395	0.973045 ± 0.0369
5 th	0.715557 ± 0.026328	0.894675 ± 0.020730	0.209965 ± 0.050093	0.963466 ± 0.0407
6 th	0.733249 ± 0.026890	0.875613 ± 0.029169	0.206690 ± 0.054850	0.990676 ± 0.0316
7 th	0.719775 ± 0.029072	0.891861 ± 0.026695	0.209265 ± 0.041337	0.966103 ± 0.0371
8 th	0.713602 ± 0.034423	0.931776 ± 0.015356	0.208923 ± 0.036814	0.973975 ± 0.0338
9 th	0.712441 ± 0.027587	0.910817 ± 0.015915	0.202283 ± 0.049044	0.971889 ± 0.0353
10 th	0.713072 ± 0.028918	0.881710 ± 0.021668	0.210118 ± 0.040435	0.991036 ± 0.0389
11 th	0.682905 ± 0.034225	0.875217 ± 0.019678	0.203195 ± 0.047049	0.949828 ± 0.0351
12 th	0.646328 ± 0.039668	0.920086 ± 0.019238	0.198986 ± 0.023012	0.960192 ± 0.0348
13 th	0.697512 ± 0.025461	0.890253 ± 0.022778	0.201406 ± 0.030348	0.960534 ± 0.0337
14 th	0.703224 ± 0.019900	0.902898 ± 0.019830	0.205575 ± 0.038635	0.932755 ± 0.0336
15 th	0.692050 ± 0.027536	0.891654 ± 0.018239	0.200860 ± 0.024085	0.929006 ± 0.0316
16 th	0.654609 ± 0.028141	0.928181 ± 0.013477	0.202355 ± 0.017180	0.963143 ± 0.0332
17 th	0.683020 ± 0.025644	0.890392 ± 0.023586	0.198464 ± 0.033714	0.941025 ± 0.0349
18 th	0.699170 ± 0.025487	0.911382 ± 0.027290	0.188802 ± 0.036537	0.950874 ± 0.0286
19 th	0.655108 ± 0.031857	0.884214 ± 0.027816	0.209234 ± 0.032466	0.962966 ± 0.0341
20 th	0.637446 ± 0.032562	0.873755 ± 0.036792	0.196807 ± 0.026001	0.952410 ± 0.0387

TABLE 5 : Résultats numériques des simulations pour chaque arrondissement avec $N = 50$ répétitions. Chaque valeur des indicateurs factice est donnée par sa moyenne sur les répétitions et la déviation standard associée. Le ratio de robustesse est calculé par rapport au premier arrondissement (choix arbitraire). Un ratio inférieur à 1 signifie que la borne de l'intégrale est plus petite pour le premier système, i.e. que l'évaluation est plus robuste pour celui-ci.

médian X_i . On définit des poids spatiaux utilisés pour quantifier l'intensité des interactions géographiques entre unités i, j , avec d_{ij} distance euclidienne entre centroïdes : $w_{ij} = \frac{P_i P_j}{(\sum_k P_k)^2} \cdot \frac{1}{d_{ij}}$ si $i \neq j$ et $w_{ii} = 0$. Les indicateurs normalisés sont les suivants

- Indice d'autocorrelation spatiale de Moran, défini comme la covariance pondérée normalisée du revenu médian par $\rho = \frac{N}{\sum_{ij} w_{ij}} \cdot \frac{\sum_{ij} w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$
- Indice de dissimilarité (proche du Moran mais intégrant les dissimilarités locales plutôt que les corrélations), donné par $d = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} |\tilde{X}_i - \tilde{X}_j|$
avec $\tilde{X}_i = \frac{X_i - \min(X_k)}{\max(X_k) - \min(X_k)}$
- Le complémentaire de l'entropie de la distribution des revenus, qui est une façon de capturer des inégalités globales $\varepsilon = 1 + \frac{1}{\log(N)} \sum_i \frac{X_i}{\sum_k X_k} \cdot \log \left(\frac{X_i}{\sum_k X_k} \right)$

De nombreuses mesures de ségrégation avec différentes signification à différentes échelles existent, comme par exemple à l'échelle d'une unité spatiale élémentaire par comparaison de la distribution de revenus empirique avec un modèle nul [LOUF et BARTHELEMY, 2015]. Le choix est ici arbitraire, afin d'illustrer la méthode avec un nombre raisonnable de dimensions.

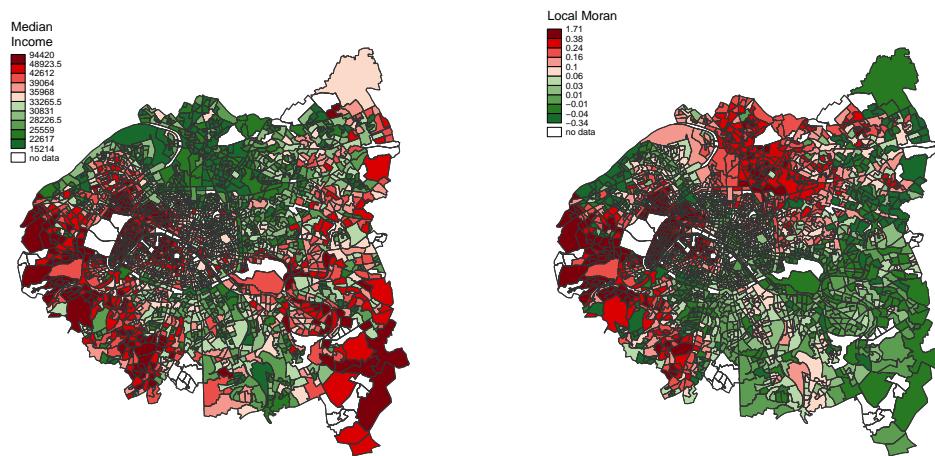


FIGURE 30 : **Cartes de ségrégation métropolitaine.** Les cartes montrent le revenu annuel médian pour les unités statistiques élémentaires (IRIS) pour les trois départements correspondant globalement à la métropole du Grand Paris, et l'index local d'autocorrelation spatiale de Moran correspondant, défini pour l'unité i par $\rho_i = N / \sum_j w_{ij} \cdot \frac{\sum_j w_{ij} (X_j - \bar{X})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$. Les zones les plus ségrégées coincident avec les plus riches et les plus pauvres, suggérant une augmentation de la ségrégation dans les cas extrêmes.

RÉSULTATS La méthode est appliquée avec ces indicateurs à la zone du Grand Paris, constitué de 4 département qui sont des ni-

veaux administratifs intermédiaires. La création récente d'un nouveau système de gouvernance métropolitaine [GILLI et OFFNER, 2009] met en évidence des interrogations sur sa pertinence, notamment sur ses capacités d'atténuer les inégalités spatiales. On peut voir en Fig. 30 les cartes de la distribution spatiale du revenu médian et de l'index local d'autocorrelation spatiale correspondant. La dichotomie bien connue entre est et ouest est retrouvée ainsi que la disparité des quartiers intra-muros, comme cela été présenté par diverses études, comme [GUÉROIS et LE GOIX, 2009] à travers l'analyse des dynamiques des transactions immobilières. Notre cadre d'étude est ensuite appliqué à une question concrète ayant des implications pour la prise de décision : *dans quelle mesure une évaluation de la ségrégation au sein de différents territoires est sensible aux données manquantes ?* Pour cela, on procède à des simulations de Monte-Carlo (75 répétitions) pour lesquelles une proportion fixe de données est supprimée aléatoirement, et l'indice de robustesse correspondant est évalué avec les indicateurs normalisés. Les simulations sont faites sur chaque département de façon indépendante, à chaque fois pour une robustesse relative à l'évaluation du Grand Paris complet. Les résultats sont présentés en Fig. 31. Toutes les zones ont une robustesse légèrement meilleure que la référence, ce qui pourrait être expliqué par une homogénéité locale et donc des indices de ségrégation plus fiables. Les implications pour la prise de décision qui peuvent être par exemple tirées sont des comparaisons directes entre les zones : une perte de 30% de l'information sur le 93 correspond à une perte de seulement 25% pour le 92. La première zone étant déjà défavorisée socio-économiquement, l'inégalité est augmentée par cette qualité moindre de l'information statistique. L'étude des déviations standard suggère des études plus approfondies comme différents régimes de réponse à la suppression de données semblent exister.

B.5.4 Discussion

Applicabilité à des situations réelles

IMPLICATIONS POUR LA PRISE DE DÉCISION L'application de notre méthode à des situations concrètes de prise de décision peut être pensée de différentes manières. Tout d'abord dans le cas d'un processus multi-attributs à but comparatif, comme la détermination d'un corridor pour une nouvelle infrastructure de transport, l'identification des territoires sur lesquels l'évaluation pourrait être biaisée (i.e. avec une mauvaise robustesse relative) devrait permettre une attention particulière pour ceux-ci, et l'adaptation des jeux de données ou la révision des points en conséquence. Dans tous les cas le processus total devrait être plus fiable. Une autre possibilité ressemble à l'application réelle que nous avons développé, i.e. la sensibilité de l'évaluation à divers paramètres comme les données manquantes. Si

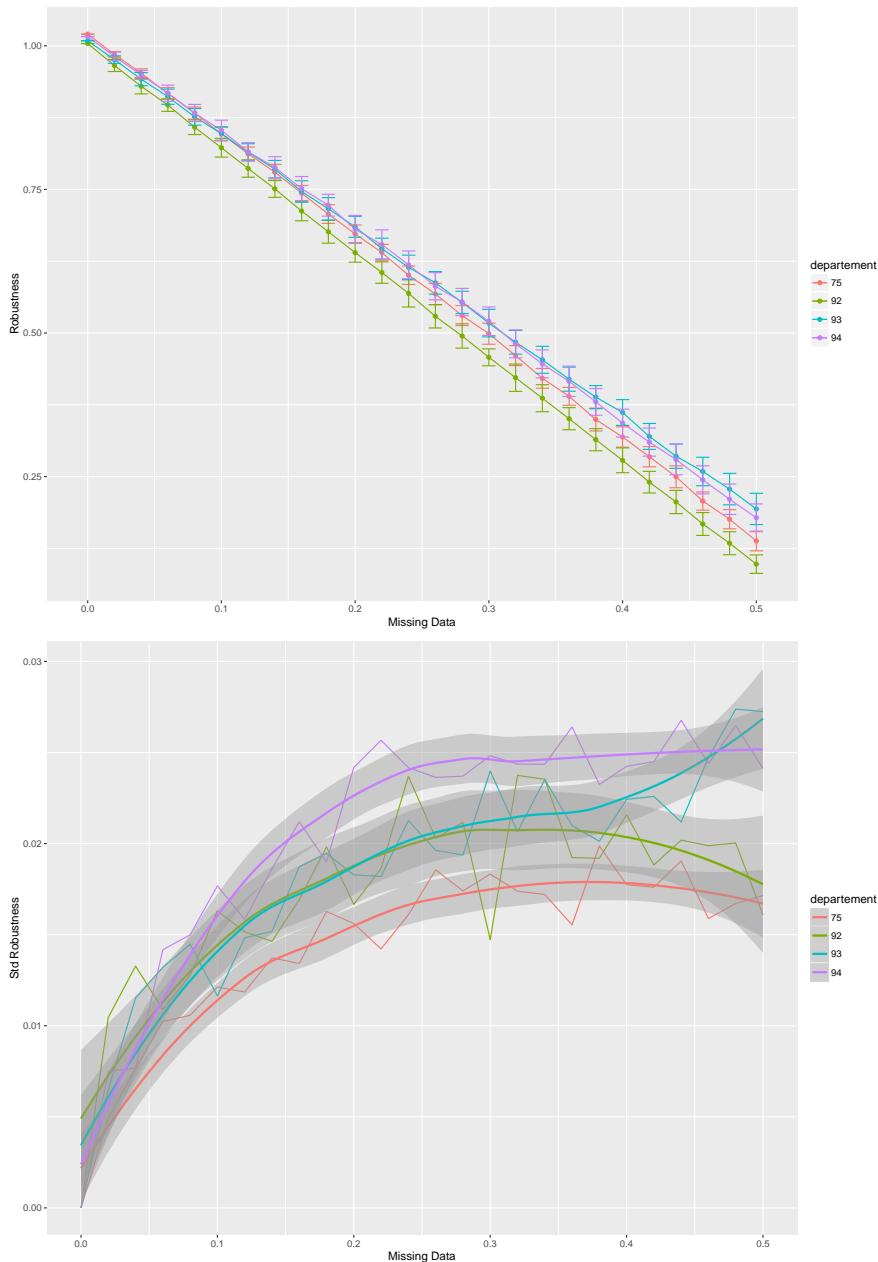


FIGURE 31 : Sensibilité de la robustesse aux données manquantes. Gauche.

Pour chaque département, des simulations de Monte-Carlo ($N=75$ répétitions) sont utilisées pour déterminer l'impact des données manquantes sur la robustesse de l'évaluation de la ségrégation. Les ratios de robustesse sont tous calculés relativement à la région métropolitaine complète avec toutes les données disponibles. Le comportement quasi-linéaire traduit une décroissance approximativement linéaire de la discrépance en fonction de la taille des données. Les trajectoires similaires des départements les plus pauvres (93,94) suggère que la correction au comportement linéaire est fonction des motifs de ségrégation. Droite. Déviations standard des ratios de robustesse. Les différents régimes (en particulier le 93 contre les autres) révèlent des transitions de phase à différents niveaux de données manquantes, signifiant que l'évaluation dans le 94 est de ce point de vue plus sensible aux données manquantes.

une décision paraît fiable car la taille de données est grande, mais que l'évaluation est très sensible à la suppression de données, il faudra être prudent pour l'interprétation des résultats et pour la prise de décision finale. Un travail approfondi et de test sera cependant nécessaire pour comprendre le comportement du cadre dans différents contextes et pouvoir piloter son application dans des situations réelles diverses.

INTÉGRATION AU SEIN DE CADRES EXISTANTS L'applicabilité de la méthode à des cas réels dépendra directement de son intégration potentielle dans des environnements existants. Au delà des difficultés techniques qui apparaissent nécessairement en essayant de coupler ou d'intégrer des implémentations existantes, des obstacles plus théoriques pourraient émerger, comme des formulations floues des fonctions ou des types de données, la cohérence des bases de données, etc. De tels cadres multi-critères sont nombreux. Un développement possible serait l'intégration dans un cadre open-source, comme par exemple celui décrit dans [TIVADAR et al., 2014] qui calcule divers indices de ségrégation urbaine, comme on l'a déjà illustré pour l'application à la ségrégation métropolitaine.

DISPONIBILITÉ DES DONNÉES BRUTES De manière générale, des données sensibles comme des questionnaires de transport, ou des données de sondage à granularité très fine, ne sont pas disponibles de manière ouverte, mais fournis de manière déjà agrégée à un certain niveau (comme par exemple les données françaises de l'Insee sont disponibles publiquement au niveau des unités statistiques élémentaires ou pour des zones plus grandes selon les variables et des contraintes de population minimale, les données plus précises étant à accès restreint). Cela signifie que l'application de notre cadre peut impliquer une procédure de recherche de données laborieuse, l'avantage d'être flexible étant alors compensé par ces contraintes additionnelles.

Validité des hypothèses théoriques

Une limitation possible de notre approche est la validité de l'hypothèse qui formule les indicateurs comme des intégrales spatiales. En fait, de nombreux indicateurs socio-économiques ne dépendent pas nécessairement directement de l'espace, et essayer de les associer à des coordonnées peut entraîner sur une pente glissante (par exemple, associer des variables économiques individuelles à des coordonnées résidentielles aura un sens seulement si la variable a une relation à l'espace, autrement un devient un artefact superflu). Même des indicateurs qui ont une valeur spatiale peuvent dériver de variables non-spatiales, comme [KWAN, 1998] le souligne au sujet de l'accèsibilité, en opposant les mesures d'accessibilité intégrée aux mesures individu-centrées mais pas forcément basée sur l'espace (comme par

exemple des décisions individuelles). Contraindre une représentation théorique d'un système pour le faire rentrer dans un cadre en changeant certaines de ses propriétés ontologiques (toujours dans le sens de la signification réelle des objets) peut être compris comme une violation d'une des règles pour la modélisation et la simulation en sciences sociales données par [BANOS, Décembre 2013], car cela impliquerait qu'il pourrait exister un langage universel pour la modélisation, malgré qu'il ne puisse retranscrire certains systèmes, ayant pour conséquences des conclusions errantes à cause d'une rupture d'ontologie dans le cas d'une formulation sur-contrainte.

Généralité du Cadre

Nous soutenons qu'un des avantages fondamentaux de notre cadre est sa généralité et sa flexibilité, puisque la robustesse des évaluations est obtenue seulement par la structure des données si l'on relaxe les hypothèses sur les valeurs des poids. Des approfondissement pourraient inclure une formulation plus générale, en supprimant par exemple l'hypothèse d'agrégation linéaire. Des fonctions d'agrégation non-linéaires demanderaient toutefois de vérifier certaines propriétés regardant les inégalités intégrales. Par exemple, des résultats similaires pourraient être obtenus en s'orientant vers des inégalités intégrales pour fonctions Lipschitziennes, comme les résultats en une dimension de [DRAGOMIR, 1999].

Conclusion

Nous avons proposé un cadre indépendant du modèle pour comparer la robustesse d'évaluations multi-attributs entre différents systèmes urbains. A partir de la discrépance des données, on fournit une définition générale de la robustesse relative sans aucune hypothèse de modèle pour le système, mais en supposant une agrégation linéaire des objectifs et des indicateurs exprimés comme des intégrales à noyaux. Nous proposons une première implémentation preuve de concept pour la ville de Paris pour laquelle les résultats numériques confirment la tendance générale attendue, et une implémentation sur des données réelles pour la ségrégation de revenus pour la région métropolitaine du Grand Paris, fournissant des réponses possibles à des questions de prise de décision plus concrètes. Des développements possibles peuvent inclure une analyse de sensibilité de la méthode, des applications à d'autres cas réels et une relaxation des hypothèses théoriques, c'est à dire de l'agrégation linéaire et de l'intégration spatiale.

C

DÉVELOPPEMENTS THÉMATIQUES

C.1 AN INTERDISCIPLINARY APPROACH TO MORPHOGENESIS

TODO : include only if very different from chapter 5

This Appendix was submitted as an Essay Paper with C. Antelope (U. California), L. Hubatsch (F. Crick Institute) and J.M. Serna (Université Paris 7), as :

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016.*

C.2 GENERATION OF CORRELATED SYNTHETIC DATA

Application : Séries temporelles financières

Contexte

Un premier domaine d'application proposé pour notre méthode est celui des séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [MANTEGNA et STANLEY, 2000] et pour lesquels les corrélations ont fait l'objet d'abondants travaux. Ainsi, l'application de la théorie des matrices aléatoires peut permettre de débruiter, ou du moins d'estimer la part de signal noyée dans le bruit, une matrice de correlations pour un grand nombre d'actifs échantillonnes à faible fréquence (retours journaliers par exemple) [BOUCHAUD et POTTERS, 2009]. De même, l'analyse de réseaux complexes construits à partir des corrélations, selon des méthodes type arbre couvrant minimal [BONANNO, LILLO et MANTEGNA, 2001] ou des extensions raffinées pour cette application précise [TUMMINELLO et al., 2005], ont permis d'obtenir des résultats prometteurs, tels la reconstruction de la structure économique des secteurs d'activités. A haute fréquence, l'estimation précise de paramètres d'interdépendance dans le cadre d'hypothèses fixées sur la dynamique, fait l'objet d'importants travaux théoriques dans un but de raffinement des modèles et des estimateurs [BARNDORFF-NIELSEN et al., 2011]. Les résultats théoriques doivent alors être testés sur des jeux de données synthétiques, qui permettent de contrôler un certain nombre de paramètres et de s'assurer qu'un effet prédit par la théorie est bien observable *toutes choses égales par ailleurs*. Par exemple, [POTIRON et MYKLAND, 2015] dérive une correction du biais de l'estimateur de *Hayashi-Yoshida* qui est un estimateur de la covariance de deux browniens corrélés à haute fréquence dans le cas de temps d'observation asynchrones, par démonstration d'un théorème de la limite centrale pour un modèle généralisé endogénisant les temps d'observations. La confirmation empirique de l'amélioration de l'estimateur est alors obtenue sur un jeu de données synthétiques à un niveau de corrélation fixé.

Formalisation

CADRE Considérons un réseau d'actifs $(X_i(t))_{1 \leq i \leq N}$ échantillonnes à haute fréquence (typiquement 1s). On se place dans un cadre multi-scalaire (utilisé par exemple dans les approches par ondelettes [RAMSEY, 2002] ou analyses multifractales du signal [BOUCHAUD, POTTERS et MEYER, 2000]) pour interpréter les signaux observés comme la superposition de composantes à des multiples échelles temporelles : $X_i = \sum_{\omega} X_i^{\omega}$. On notera $T_i^{\omega} = \sum_{\omega' \leq \omega} X_i^{\omega'}$ le signal filtré à une fréquence ω donnée. Prédire l'évolution d'une composante à une échelle donnée est alors un problème caractéristique de l'étude des

systèmes complexes, pour lequel l'enjeu est l'identification de régularités et leur distinction des composantes considérées comme stochastiques en comparaison¹. Dans un souci de simplicité, on représente un tel processus par un modèle de prédiction de tendance à une échelle temporelle ω_1 donnée, formellement un estimateur $M_{\omega_1} : (T_i^{\omega_1}(t'))_{t' < t} \mapsto \hat{T}_i^{\omega_1}(t)$ dont l'objectif est la minimisation de l'erreur sur la tendance réelle $\|T_i^{\omega_1} - \hat{T}_i^{\omega_1}\|$. Dans le cas d'estimateurs auto-regressifs multivariés, la performance dépendra entre autre des correlations respectives entre actifs et il est alors intéressant d'utiliser la méthode pour évaluer celle-ci en fonction de niveaux de correlation à plusieurs échelles. On assume une dynamique de Black-Scholes [JARROW, 1999] pour les actifs, i.e. $dX = \sigma \cdot dW$ avec W processus de Wiener, ce qui permettra d'obtenir facilement des niveaux de correlation voulus.

GÉNÉRATION DES DONNÉES Il est alors aisé de générer \tilde{X}_i tel que $\text{Var}[\tilde{X}_i^{\omega_1}] = \Sigma R$ (Σ variance estimée et R matrice de corrélation fixée), par la simulation de processus de Wiener au niveau de corrélation fixé et tel que $X_i^{\omega \leq \omega_0} = \tilde{X}_i^{\omega \leq \omega_0}$ (critère de proximité au données : les composantes à plus basse fréquence qu'une fréquence fondamentale $\omega_0 < \omega_1$ sont identiques). En effet, si $dW_1 \perp\!\!\!\perp dW_1^{\perp\!\!\!\perp}$ (et $\sigma_1 < \sigma_2$ pour fixer les idées, quitte à échanger les actifs), alors $W_2 = \rho_{12}W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} W_1^{\perp\!\!\!\perp}$ est tel que $\rho(dW_1, dW_2) = \rho_{12}$. Les signaux suivants sont construits de la même manière par ortho-normalisation de Gram. On isole alors la composante à la fréquence ω_1 voulue par filtrage, c'est à dire $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$ (avec \mathcal{F}_{ω_0} filtre passe-bas à fréquence de coupure ω_0), puis on reconstruit les signaux synthétiques par $\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1}$.

Implémentation et résultats

MÉTHODOLOGIE La méthode est testée sur un exemple de deux actifs du marché des devises (EUR/USD et EUR/GBP), sur une période de 6 mois de juin 2015 à novembre 2015. Le nettoyage des données², originellement échantillonnées à l'ordre de la seconde, consiste dans un premier temps à la détermination du support temporel commun maximal (les séquences manquantes étant alors ignorées, par translation verticale des séries, i.e. $S(t) := S(t) \cdot \frac{S(t_n)}{S(t_{n-1})}$ lorsque t_{n-1}, t_n sont les extrémités du "trou" et $S(t)$ la valeur de l'actif, ce qui revient à garder la contrainte d'avoir des retours à pas de temps simi-

¹ voir [GELL-MANN, 1995] pour une discussion étendue sur la construction de *schema* pour l'étude de systèmes complexes adaptatifs (par des systèmes complexes adaptatifs).

² obtenues depuis <http://www.histdata.com/>, sans licence spécifiée, les données nettoyées et filtrées à ω_m uniquement sont mises en accessibilité pour respect du copyright.

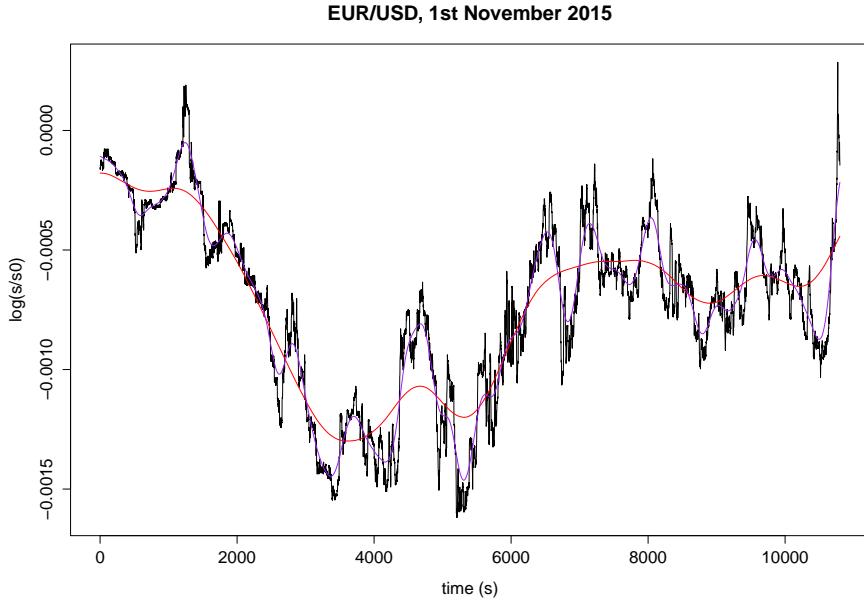


FIGURE 32 :

liaires entre actifs). On étudie alors les *log-prix* et *log-retours*, définis par $X(t) := \log \frac{S(t)}{S_0}$ et $\Delta X(t) = X(t) - X(t-1)$. Les données brutes sont filtrées à une fréquence $\omega_m = 10\text{min}$ (qui sera la fréquence maximale d'étude) pour un souci de performance computationnelle. On utilise un filtre gaussien non causal de largeur totale ω . On fixe $\omega_0 = 24\text{h}$ et on se propose de construire des données synthétiques aux fréquences $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. Voir la figure 32 pour un exemple de la structure du signal à ces différentes échelles.

Il est crucial de noter l'interférence entre les fréquences ω_0 et ω_1 dans le signal construit : la corrélation effectivement estimée est

$$\rho_e = \rho [\Delta \tilde{X}_1, \Delta \tilde{X}_2] = \rho [\Delta T_1^{\omega_0} + \Delta \tilde{X}_1^\omega, \Delta T_2^{\omega_0} + \Delta \tilde{X}_2^\omega]$$

ce qui conduit à dériver dans la limite raisonnable $\sigma_1 \gg \sigma_0$ (fréquence fondamentale suffisamment basse), lorsque $\text{Cov} [\Delta \tilde{X}_i^{\omega_1}, \Delta X_j^\omega] = 0$ pour tous $i, j, \omega_1 > \omega$, et les retours d'espérance nulle à toutes échelles, en notant $\rho_0 = \rho [\Delta T_1^{\omega_0}, \Delta T_2^{\omega_0}]$, $\rho = \rho [\tilde{X}_1^{\omega_1}, \tilde{X}_2^{\omega_1}]$, et $\varepsilon_i = \frac{\sigma(\Delta T_i^{\omega_0})}{\sigma(\Delta \tilde{X}_i^{\omega_1})}$, la correction sur la corrélation effective due aux interférences : la corrélation effective est alors au premier ordre

$$\rho_e = [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right] \quad (19)$$

ce qui donne l'expression de la corrélation que l'on pourra effectivement simuler dans les données synthétiques.

La corrélation est estimée par méthode de Pearson, avec l'estimateur de la covariance au biais corrigé, c'est à dire $\hat{\rho}[X_1, X_2] = \frac{\hat{C}[X_1, X_2]}{\sqrt{\text{Var}[X_1]\text{Var}[X_2]}}$, où $\hat{C}[X_1, X_2] = \frac{1}{(T-1)} \sum_t X_1(t)X_2(t) - \frac{1}{T(T-1)} \sum_t X_1(t) \sum_t X_2(t)$ et $\text{Var}[X] = \frac{1}{T} \sum_t X^2(t) - (\frac{1}{T} \sum_t X(t))^2$.

Le modèle de prédiction M_{ω_1} testé est simplement un modèle ARMA pour lequel on fixe les paramètres $p = 2, q = 0$ (on ne crée pas de corrélation retardée, on ne s'attend donc pas à de grand ordre d'auto-regression, les signaux originaux étant à mémoire relativement courte ; de plus le lissage n'est pas nécessaire puisqu'on travaille sur des données filtrées), appliqué de manière adaptative³. Plus précisément, étant donné une fenêtre temporelle T_W , on estime pour tout t le modèle sur $[t - T_W + 1, t]$ afin de prédire les signaux à $t + 1$.

IMPLÉMENTATION L'implémentation est faite en langage R, utilisant en particulier la bibliothèque MTS [Tsay, 2015] pour les modèles de séries temporelles. Les données nettoyées et le code source sont disponibles de manière ouverte sur le dépôt git du projet⁴.

RÉSULTATS La figure ?? donne les corrélations effectives calculées sur les données synthétiques. Pour des valeurs standard des paramètres (par exemple pour $\omega_0 = 24h$, $\omega_1 = 2h$ et $\rho = -0.5$), on a $\rho_0 \simeq 0.71$ et $\varepsilon_i \simeq 0.3$ et ainsi $|\rho_e - \rho| \simeq 0.05$. On constate dans l'intervalle $\rho \in [-0.5, 0.5]$ un bon accord entre la valeur ρ_e prédictive par 19 et les valeurs observées, et une déviation pour de plus grandes valeurs absolues, d'autant plus grande que ω_1 est petit : cela confirme l'intuition que lorsque la fréquence descend et se rapproche de ω_0 , les interférences entre les deux composantes vont devenir non négligeables et invalider les hypothèses d'indépendance par exemple.

On applique ensuite le modèle prédictif décrit ci-dessus aux données synthétiques, afin d'étudier sa performance moyenne en fonction du niveau de corrélation des données. Les résultats pour $\omega_1 = 1h, 1h30, 2h$ sont présentés en figure 34. Le résultat a priori contre-intuitif d'une performance maximale à corrélation nulle pour l'un des actifs confirme l'intérêt d'une génération de données hybrides : l'étude des corrélations décalées (*lagged correlations*) montre une dissymétrie présente dans les données réelles, interprété à l'échelle journalière comme une influence augmentée de EURGBP sur EURUSD à 2h de décalage environ. L'existence de ce *lag* permet une "bonne"

³ il s'agit d'un niveau d'adaptation relativement faible, les paramètres T_W, p, q et même le type de modèle restant fixés. On se place ainsi dans le cadre de [POTIRON, 2016] qui suppose une dynamique localement paramétrique, mais pour lequel on fixe les méta-paramètres de la dynamique. On pourrait imaginer estimer un T_W variable qui s'adapterait pour une meilleure estimation locale, à l'image de l'estimation de paramètres en traitement du signal Bayesien effectuée via augmentation de l'état par les paramètres.

⁴ at <https://github.com/JusteRaimbault/SynthAsset>

prédition de EURUSD due à la fréquence fondamentale, perturbée par le bruit ajouté, de façon proportionnelle à sa correlation : plus les bruits sont corrélés, plus le modèle les prendra en compte et se trompera plus à cause du caractère markovien des browniens simulés⁵.

L'exemple présenté ici est un *modèle jouet* et n'a pas d'application pratique, mais démontre l'intérêt de l'utilisation des données synthétiques simulées. On peut imaginer simuler des données plus proches de la réalité (existence de motifs réalistes de *lagged correlation* par exemple, modèles plus réalistes que le Black-Scholes) et appliquer la méthode sur des modèles plus opérationnels.

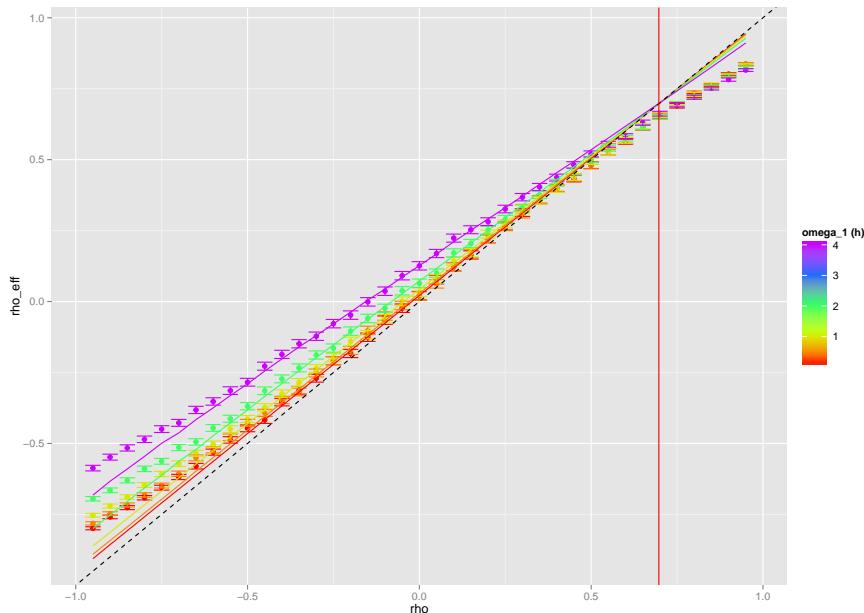


FIGURE 33 :

⁵ en théorie le modèle utilisé n'a aucun pouvoir prédictif sur des browniens purs

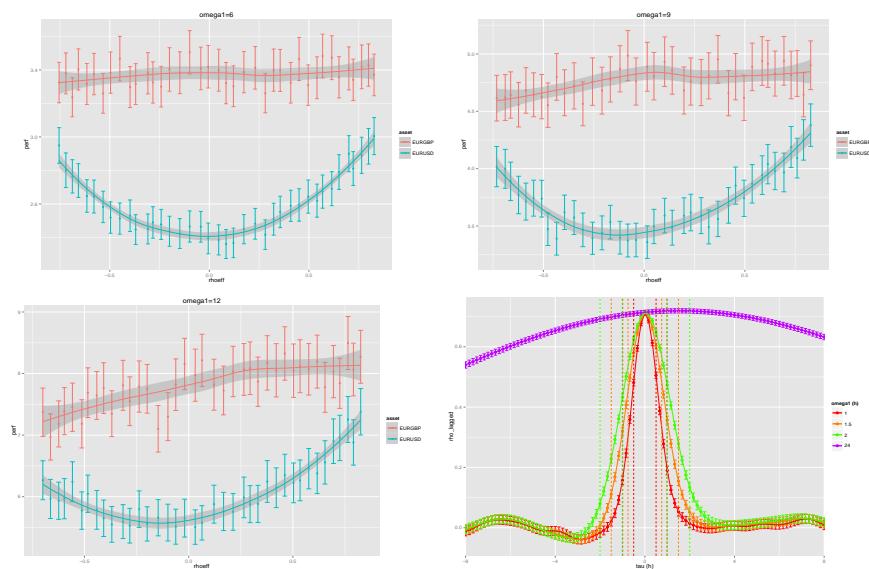


FIGURE 34 :

C.3 CLASSIFYING PATENTS BASED ON THEIR SEMANTIC CONTENT

In this paper, we extend some usual techniques of classification resulting from a large-scale data-mining and network approach. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract the relevant keywords accordingly. We refer to this classification as *semantic approach* in contrast with the more common *technological approach* which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.

Introduction

Innovation and technological change have been described by many scholars as the main drivers of economic growth as in [AGHION et HOWITT, 1992] and [ROMER, 1990]. [GRILICHES, 1990] advertised the use of patents as an economic indicator and as a good proxy for innovation. Subsequently, the easier availability of comprehensive databases on patent details and the increasing number of studies allowing a more efficient use of these data (e.g. [HALL, JAFFE et TRAJTENBERG, 2001]) have opened the way to a very wide range of analysis. Most of the statistics derived from the patent databases relied on a few key features : the identity of the inventor, the type and identity of the rights owner, the citations made by the patent to prior art and the technological classes assigned by the patent office post patent's content review. Combining this information is particularly relevant when trying to capture the diffusion of knowledge and the interaction between technological fields as studied in [YOUN et al., 2015]. With methods such as citation dynamics modeling discussed in [NEWMAN, 2013] or co-authorship networks analysis in [SARIGÖL et al., 2014], a large body of the literature such as [SORENSEN, RIVKIN et FLEMING, 2006] or [KAY et al., 2014] has studied patents citation network to understand processes driving technological innovation, diffusion and the birth of technological clusters. Finally, [BRUCK et al., 2016] look at the dynamics of citations from different classes to show that the laser/ink-jet printer technology resulted from the recombination of two different existing technologies.

Consequently, technological classification combined with other features of patents can be a valuable tool for researchers interested in studying technologies throughout history and to predict future inno-

vations by looking at past knowledge and interaction across sectors and technologies. But it is also crucial for firms that face an ever changing demand structure and need to anticipate future technological trends and convergence (see, e.g., [CURRAN et LEKER, 2011]) to adapt to the resulting increase in competition discussed in [KATZ, 1996] and to maintain market share. Curiously, and in spite of the large number of studies that analyze interactions across technologies [FURMAN et STERN, 2011], little is known about the underlying “innovation network” (e.g. [ACEMOGLU et KERR, 2016]).

In this monograph, we propose an alternative classification based on semantic network analysis from patent abstracts and explore the new information emerging from it. In contrast with the regular technological classification which results from the choice of the patent reviewer, semantic classification is carried automatically based on the content of the patent abstract. Although patent officers are experts in their fields, the relevance of the existing classification is limited by the fact that it is based on the state of technology at the time the patent was granted and cannot anticipate the birth of new fields. To correct for this, the USPTO regularly make changes in its classification in order to adapt to technological change (for example, the “nanotechnology” class (977) was established in 2004 and retroactively to all relevant previously granted patents). In contrast we don’t face this issue with the semantic approach. The semantic links can be clues of one technology taking inspiration from another and good predictors of future technology convergence (e.g. [PRESCHITSCHER et al., 2013] study semantic similarities from the whole text of 326 US-patents on *phytosterols* and show that semantic analysis have a good predicting power of future technology convergence). One can for instance consider the case of the word *optic*. Until more recently, this word was often associated with technologies such as photography or eye surgery, while it is now almost exclusively used in a context of semi-transistor design and electro-optic. This semantic shift did not happen by chance but contains information on the fact that modern electronic extensively uses technologies that were initially developed in optic.

Previous research has already proposed to use semantic networks to study technological domains and detect novelty. [YOUN et PARK, 2004] was one of the first to enhance this approach with the idea of visualizing keywords network illustrated on a small technological domain. The same approach can be used to help companies identifying the state of the art in their field and avoid patent infringement as in [PARK et YOUN, 2014] and [YOUN et KIM, 2011]. More closely related to our methodology, [GERKEN et MOEHRLE, 2012] develop a method based on patent semantic analysis of patent to vindicate the view that this approach outperform others in the monitoring of technology and in the identification of novelty innovation. Semantic analysis has al-

ready proven its efficiency in various fields, such as in technology studies (e.g. [CHOI et HWANG, 2014] and [FATTORI, PEDRAZZI et TURRA, 2003]) and in political science (e.g. [GURCIULLO et al., 2015]).

Building on such previous research, we make several contributions by fulfilling some shortcomings of existing studies, such as for example the use of frequency-selected single keywords. First of all, we develop and implement a novel fully-automatized methodology to classify patents according to their semantic abstract content, which is to the best of our knowledge the first of its type. This includes the following refinements for which details can be found in Section ?? : (i) use of multi-stems as potential keywords; (ii) filtering of keywords based on a second-order (co-occurrences) relevance measure and on an external independent measure (technological dispersion); (iii) multi-objective optimization of semantic network modularity and size. The use of all this techniques in the context of semantic classification is new and essential from a practical perspective.

Furthermore, most of the existing studies rely on a subsample of patent data, whereas we implement it on the full US Patent database from 1976 to 2013. This way, a general structure of technological innovation can be studied. We draw from this application promising qualitative stylized facts, such as a qualitative regime shift around the end of the 1990s, and a significant improvement of citation modularity for the semantic classification when comparing to the technological classification. These thematic conclusions validate our method as a useful tool to extract endogenous information, in a complementary way to the technological classification.

Finally, the statistical model introduced in Section seems to indicate that patents tend to cite more similar patents in the semantic network when fitted to data. In particular, this propensity is shown to be significantly bigger than the corresponding propensity for technological classes, and this seems to be consistent over time. On the account of this information, we believe that patent officers could benefit very much from looking at the semantic network when considering potential citation candidates of a patent in review.

The paper is organized as follows. Section ?? presents the patent data, the existing classification and provide details about the data collection process. Section ?? explains the construction of the semantic classes. Section ?? tests their relevance by providing exploratory results. Finally, section ?? discusses potential further developments and conclude. More details, including robustness checking, figures and technical derivations can be found in ??, ?? and ??.

Background

In our analysis, we will consider all utility patents granted in the United States Patent and Trademark Office (USPTO) from 1976 to 2013.

A clearer definition of utility patent is given in [??](#). Also, additional information on how to correctly exploit patent data can be found in [HALL, JAFFE et TRAJTENBERG, [2001](#)] and [LERNER et SERU, [2015](#)].

An existing classification : the USPC system

Each USPTO patent is associated with a non-empty set of technological classes and subclasses. There are currently around 440 classes and over 150,000 subclasses constituting the United State Patent Classification (USPC) system. While a technological class corresponds to the technological field covered by the patent, a subclass stands for a specific technology or method used in this invention. A patent can have multiple technological classes, on average in our data a patent has 1.8 different classes and 3.9 pairs of class/subclass. At this stage, two features of this system are worth mentioning : (i) classes and subclasses are not chosen by the inventors of the patent but by the examiner during the granting process based on the content of the patent; (ii) the classification has evolved in time and continues to change in order to adapt to new technologies by creating or editing classes. When a change occurs, the USPTO reviews all the previous patents so as to create a consistent classification.

A bibliographical network between patents : citations

As with scientific publications, patents must give reference to all the previous patents which correspond to related prior art. They therefore indicate the past knowledge which relates to the patented invention. Yet, contrary to scientific citations, they also have an important legal role as they are used to delimit the scope of the property rights awarded by the patent. One can consult [OECD, [2009](#)] for more details about this. Failing to refer to prior art can lead to the invalidation of the patent (e.g. [DECHEZLEPRÂTRE, MARTIN et MOHNEN, [2014](#)]). Another crucial difference is that the majority of the citations are actually chosen by the examiners and not by the inventors themselves. From the USPTO, we gather information of all citations made by each patent (backward citations) and all citations received by each patent as of the end of 2013 (forward citations). We can thus build a complete network of citations that we will use later on in the analysis.

Turning to the structure of the lag between the citing and the cited patent in terms of application date, we see that the mean of this lag is 8.5 years and the median is 7 years. This distribution is highly skewed, the 95th percentile is 21 years. We also report 164,000 citations with a negative time lag. This is due to the fact that some citations can be added during the examination process and some patents require more time to be granted than others.

In what follows, we choose to restrict attention to pairs of citations with a lag no larger than 5 years. We impose this restriction for two

reasons. First, the number of citations received peaks 4-5 years after application. Second, the structure of the citation lag is necessarily biased by the truncation of our sample : the more recent patents mechanically receive less citations than the older ones. As we are restricting to citations received no later than 5 years after the application date, this effect will only affect patents with an application date after 2007.

Data collection and basic description

Each patent contains an abstract and a core text which describe the invention. To see what a patent looks like in practice, one can refer to the USPTO patent full-text database <http://patft.uspto.gov/netahtml/PTO/index.html> or to Google patent which publishes USPTO patents in pdf format at <https://patents.google.com>. Although including the full core texts would be natural and probably very useful in a systematic text-mining approach as done in [TSENG, LIN et LIN, 2007], they are too long to be included and thus we consider only the abstracts for the analysis. Indeed, the semantic analysis counts more than 4 million patents, with corresponding abstracts with an average length of 120.8 words (and a standard deviation of 62.4), a size that is already challenging in terms of computational burden and data size. In addition, abstracts are aimed at synthesizing purpose and content of patents and must therefore be a relevant object of study (see [ADAMS, 2010]). The USPTO defines a guidance stating that an abstract should be “a summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention” (PCT Rule 8).

We construct from raw data a unified database. Data is collected from USPTO patent redbook bulk downloads, that provides as raw data (specific dat or xml formats) full patent information, starting from 1976. Detailed procedure of data collection, parsing and consolidation are available in ?? . The latest dump of the database in Mongodb format is available at <http://dx.doi.org/10.7910/DVN/BW3ACK>. Collection and homogenization of the database into a directly usable database with basic information and abstracts was an important task as USPTO raw data formats are involved and change frequently.

We count 4,666,365 utility patents with an abstract granted from 1976 to 2013. A very small number of patents have a missing abstract, these are patents that have been withdrawn and we do not consider them in the analysis. The number of patents granted each year increases from around 70,000 in 1976 to about 278,000 in 2013. When distributed by the year of application, the picture is slightly different. The number of patents steadily increase from 1976 to 2000 and re-

mains constant around 200,000 per year from 2000 to 2007. Restricting our sample to patent with application date ranging from 1976 to 2007, we are left with 3,949,615 patents. These patents cite 38,756,292 other patents with the empirical lag distribution that has been extensively analyzed in [HALL, JAFFE et TRAJTENBERG, 2001]. Conditioned on being cited at least once, a patent receives on average 13.5 citations within a five-year window. 270,877 patents receive no citation during the next five years following application, 10% of patents receive only one citation and 1% of them receive more than 100 citations. A within class citation is defined as a citation between two patents sharing at least one common technological class. Following this definition, 84% of the citations are within class citations. 14% of the citations are between two patents that share the exact same set of technological classes.

Towards a Complementary Classification

Potentialities of text-mining techniques as an alternative way to analyze and classify patents are documented in [TSENG, LIN et LIN, 2007]. The author's main argument, in support of an automatic classification tool for patent, is to reduce the considerable amount of human effort needed to classify all the applications. The work conducted in the field of natural language processing and/or text analysis has been developed in order to improve search performance in patent databases, build technology map or investigate the potential infringement risks prior to developing a new technology (see [ABBAS, ZHANG et KHAN, 2014] for a review). Text-mining of patent documents is also widely used as a tool to build networks which carry additional information to the simplistic bibliographic connections model as argued in [YOUN et PARK, 2004]. As far as the authors know, the use of text-mining as a way to build a global classification of patents remains however largely unexplored. One notable exception can be found in [PRESCHITSCHET AL., 2013] where semantic-based classification is shown to outperform the standard classification in predicting the convergence of technologies even in small samples. Semantic analysis reveals itself to be more flexible and more quickly adaptable to the apparition of new clusters of technologies. Indeed, as argued in [PRESCHITSCHET AL., 2013], before two distinct technologies start to clearly converge, one should expect similar words to be used in patents from both technologies.

Finally, a semantic classification where patents are gathered based on the fact that they share similar significant keywords has the advantage of including a network feature that cannot be found in the USPC case, namely that each patent is associated with a vector of probability to belong to each of the semantic classes (more details on this feature can be found in Section C.3). Using co-occurrence of keywords, it is then possible to construct a network of patents and

to study the influence of some key topological features. As reviewed previously, the use of co-occurrences is the usual way to construct a semantic network. Other hybrid technique such as bipartite semantic/authors networks, do not have the nice feature of relying solely on endogenous semantic information contained in data.

Semantic Classification Construction

In this section, we describe methods and empirical analysis leading to the construction of semantic network and the corresponding classification.

Keywords extraction

Let \mathcal{P} be the set of patents, we first assign to a patent $p \in \mathcal{P}$ a set of potentially significant keywords $K(p)$ from its text $A(p)$ (that corresponds to the concatenation of its own title and abstract). $K(p)$ are extracted through a similar procedure as the one detailed in [CHAVALARIAS et COINTET, 2013] :

1. Text parsing and Tokenization : we transform raw texts into a set of words and sentences, reading it (parsing) and splitting it into elementary entities (words organized in sentences).
2. Part-of-speech tagging : attribution of a grammatical function to each of the tokens defined previously.
3. Stem extraction : families of words are generally derived from a unique root called stem (for example compute, computer, computation all yield the same stem comput) that we extract from tokens. At this point the abstract text is reduced to a set of stems and their grammatical functions.
4. Multi-stems construction : these are the basic semantic units used in further analysis. They are constructed as groups of successive stems in a sentence which satisfies a simple grammatical function rule. The length of the group is between 1 and 3 and its elements are either nouns, attributive verbs or adjectives. We choose to extract the semantics from such nominal groups in view of the technical nature of texts, which is not likely to contain subtle nuances in combinations of verbs and nominal groups.

Text processing operations are implemented in python in order to use built-in functions `nltk` library [NLTK, 2015] for most of above operations. This library supports most of state-of-the-art natural language processing operations. Source code is openly available on the repository of the project at <https://github.com/JusteRaimbault/PatentsMining>.

Keywords relevance estimation

RELEVANCE DEFINITION Following the heuristic in [CHAVALARIAS et COINTET, 2013], we estimate relevance score in order to filter multi-stem. The choice of the total number of keywords to be extracted, which we shall denote K_w , is important, too small a value would yield similar network structures but including less information whereas very large values tend to include too many irrelevant keywords. We choose to set this parameter to $K_w = 100,000$. We first consider the filtration of $k \cdot K_w$ (with $k = 4$) to keep a large set of potential keywords but still have a reasonable number of co-occurrences to be computed. This step has only very marginal effects on the nature of the final keywords but is necessary for computational purposes. The filtration is done on the *unithood* u_i , defined for keyword i as $u_i = f_i \cdot \log(1 + l_i)$ where f_i is the multi-stem's number of apparitions over the whole corpus and l_i its length in words. A second filtration of K_w keywords is done on the *termhood* t_i , where the formal definition can be found in (20). It is computed as a chi-squared score on the distribution of the stem's co-occurrences and then compared to a uniform distribution within the whole corpus. Intuitively, uniformly distributed terms will be identified as plain language and they are thus not relevant for the classification. More precisely, we compute the co-occurrence matrix (M_{ij}) , where M_{ij} is defined as the number of patents where stems i and j appear together. The *termhood* score t_i is defined as

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}. \quad (20)$$

MOVING WINDOW ESTIMATION The previous scores are estimated on a moving window with fixed time length following the idea that the present relevance is given by the most recent context and thus that the influence vanishes when going further into the past. Consequently, the co-occurrence matrix is chosen to be constructed at year t restricting to patent which applied during the time window $[t - T_0; t]$. Note that the causal property of the window is crucial as the future cannot play any role in the current state of keywords and patents. This way, we will obtain semantic classes which are exploitable on a T_0 time span. For example, this enables us to compute the modularity of classes in the citation network as in section C.3. In the following, we take $T_0 = 4$ (which corresponds to a five year window) consistently with the choice of maximum time lag for citations made in Section ???. Accordingly, the sensitivity analysis for $T_0 = 2$ can be found in Appendix ??.

Construction of the semantic network

We keep the set of most relevant keywords \mathcal{K}_W and obtain their co-occurrence matrix as defined in Section C.3. This matrix can be directly interpreted as the weighted adjacency matrix of the semantic network. At this stage, the topology of raw networks does not allow the extraction of clear communities. This is partly due to the presence of hubs that correspond to frequent terms common to many fields (e.g. method, apparat) which are wrongly filtered as relevant. We therefore introduce an additional measure to correct the network topology : the concentration of keywords across technological classes, defined as :

$$c_{\text{tech}}(s) = \sum_{j=1}^{N^{(\text{tec})}} \frac{k_j(s)^2}{(\sum_i k_i(s))^2},$$

where $k_j(s)$ is the number of occurrences of the s th keyword in each of the j th technological class taken from one of the $N^{(\text{tec})}$ USPC classes. The higher c_{tech} , the more specific to a technological class the node is. For example, the terms **semiconductor** is widely used in electronics and does not contain any significant information in this field. We use a threshold parameter, defined as θ_c , and keep nodes with $c_{\text{tech}}(s) > \theta_c$. Likewise, edges with low weights correspond to rare co-occurrences and are considered to be noise. To account for this we define the threshold parameter for edges θ_w , and we filter edges with a weight below θ_w , following the rationale that two keywords are not linked “by chance” if they appear simultaneously a minimal number of time. To control for size effect, we normalize by taking $\theta_w = \theta_w^{(0)} \cdot N_p$ where N_p is the number of patents in the corpus ($N_p = |\mathcal{P}|$). $\theta_w^{(0)}$ is thus a varying parameter interpreted as a noise threshold *per patent*. Communities are then extracted using a standard modularity maximization procedure as described in [CLAUSET, NEWMAN et MOORE, 2004] to which we add the two constraints captured by θ_w and θ_c , namely that edges must have a weight greater than θ_w and nodes a concentration greater than θ_c . At this stage, both parameters θ_c and $\theta_w^{(0)}$ are unconstrained and their choice is not straightforward. Indeed, many optimization objectives are possible, such as the modularity, network size or number of communities. We find that modularity is maximized at a roughly stable value of θ_w across different θ_c for each year, corresponding to a stable $\theta_w^{(0)}$ across years, which leads us to choose $\theta_w^{(0)} = 4.1 \cdot 10^{-5}$. Then for the choice of θ_c , different candidates points lie on a Pareto front for the bi-objective optimization on number of communities and network size. There is a priori no reason to choose any specific point among the different optima. Consequently, we have tried the analysis with all the candidate values for θ_c and found that the results are the most

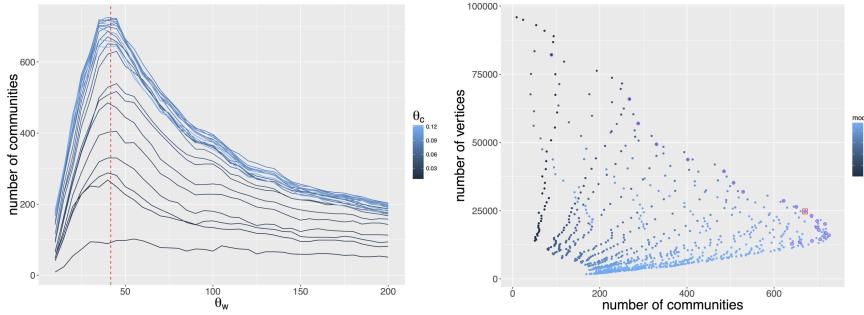


FIGURE 35 : Sensitivity analysis of network community structure to filtering parameters. We consider a specific window 2000-2004 and the obtained plots are typical. (*Left panel*) We plot the number of communities as a function of the edge threshold parameter θ_w for different values of the node threshold parameter θ_c . The maximum is roughly stable across θ_c (dashed red line). (*Right panel*) To choose θ_c , we do a Pareto optimization on communities and network size : the compromise point (red overline) on the Pareto front (purple overline : possible choices after having fixed $\theta_w^{(0)}$; blue level gives modularity) corresponds to $\theta_c = 0.06$.

reasonable when taking $\theta_c = 0.06$ (see Fig. 35). We show in Fig. 36 an example of semantic network visualization.

Characteristics of Semantic Classes

For each year t , we define as $N_t^{(sem)}$ the number of semantic classes which have been computed by clustering keywords from patents appeared during the period $[t - T_0, t]$ (we recall that we have chosen $T_0 = 4$). Each semantic class $k = 1, \dots, N_t^{(sem)}$ is characterized by a set of keywords $K(k, t)$ which is a subset of \mathcal{K}_W selected as described in previous sections. The cardinal of $K(k, t)$ distribution across each semantic class k is highly skewed with a few semantic classes containing over 1,000 keywords, most of them with roughly the same number of keywords. In contrast, there are also many semantic classes with only two keywords. There are around 30 keywords by semantic class on average and the median is 2 for any t . Fig. 37 shows that the average number of keywords is relatively stable from 1976 to 1992 and then picks around 1996 prior to going down.

TITLE OF SEMANTIC CLASSES USPC technological classes are defined by a title and a highly accurate definition which help retrieve patents easily. The title can be a single word (e.g. : class 101 : "Printing") or more complex (e.g. : class 218 : "High-voltage switches with arc preventing or extinguishing devices"). As our goal is to release a comprehensive database in which each patent is associated with a set of semantic classes, it is necessary to give an insight on what

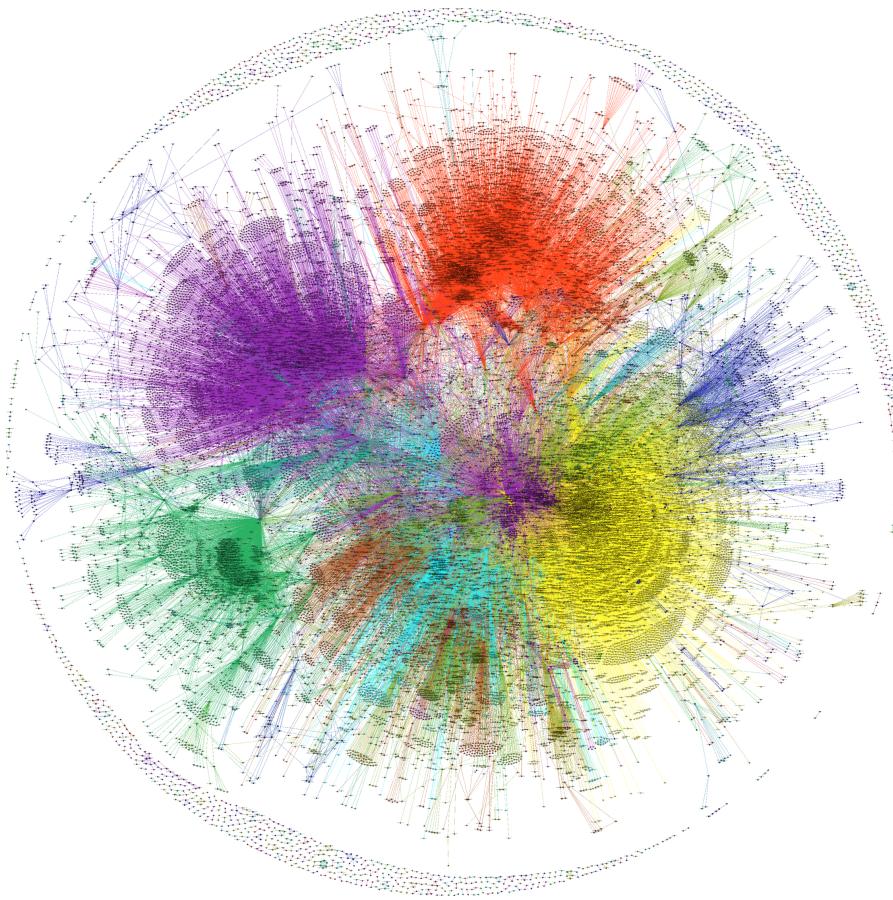


FIGURE 36 : An example of semantic network visualization. We show the network obtained for the window 2000-2004, with parameters $\theta_c = 0.06$ and $\theta_w = \theta_w^{(0)} \cdot N_p = 4.5e^{-5} \cdot 9.1e^5$. The corresponding file in a vector format (.svg), that can be zoomed and explored, is available as ??.

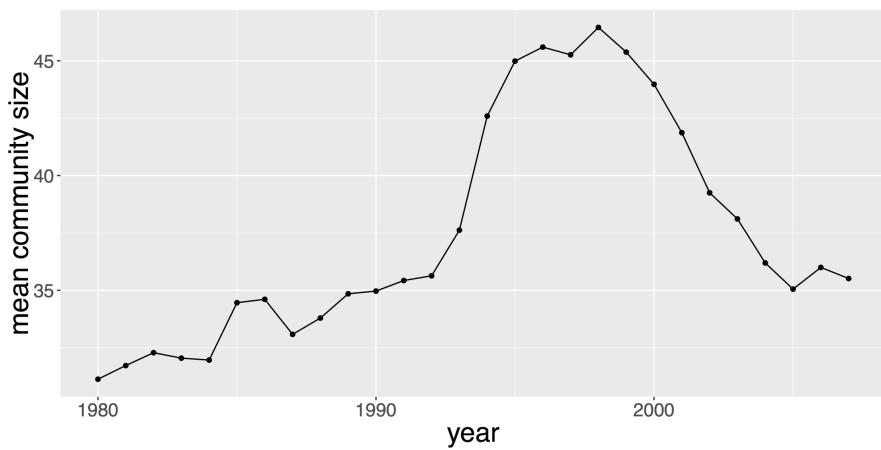


FIGURE 37 : This figure plots the average number of keywords by semantic class for each time window $[t-4; t]$ from $t = 1980$ to $t = 2007$.

these classes represent by associating a short description or a title as in [TSENG, LIN et LIN, 2007]. In our case, such description is taken as a subset of keywords taken from $K(k, t)$. For the vast majority of semantic classes that have less than 5 keywords, we decide to keep all of these keywords as a description. For the remaining classes which feature around 50 keywords on average, we rely on the topological properties of the semantic network. [YANG et al., 2000] suggest to retain only the most frequently used terms in $K(k, t)$. Another possibility is to select 5 keywords based on their network centrality with the idea that very central keywords are the best candidates to describe the overall idea captured by a community. For example, the largest semantic class in 2003-2007 is characterized by the keywords : Support Packet; Tree Network; Network Wide; Voic Stream; Code Symbol Reader.

SIZE OF TECHNOLOGICAL AND SEMANTIC CLASSES We consider a specific window of observations (for example 2000-2004), and we define Z the number of patents which appeared during that time window. For each patent $i = 1, \dots, Z$ we associate a vector of probability where each component $p_{ij}^{(sem)} \in [0, 1]$, with $j = 1, \dots, N(sem)$ and where

$$\sum_{j=1}^{N(sem)} p_{ij}^{(sem)} = 1$$

(when there is no room for confusion, we drop the subscript t in $N_t^{(sem)}$). On average across all time windows, a patent is associated to 1.8 semantic classes with a positive probability. Next we define the size of a semantic class as

$$S_j^{(sem)} = \sum_{i=1}^Z p_{ij}^{(sem)}.$$

Correspondingly, we aim to provide a consistent definition for technological classes. For that purpose, we follow the so-called “fractional count” method, which was introduced by the USPTO and consists in dividing equally the patents between all the classes they belong to. Formally, we define the number of technological classes as $N^{(tec)}$ (which is not time dependent contrary to the semantic case) and for $j = 1, \dots, N^{(tec)}$ the corresponding matrix of probability is defined as

$$p_{ij}^{(tec)} = \frac{B_{ij}}{\sum_{k=1}^{N^{(tec)}} B_{ik}},$$

where B_{ij} equals 1 if the i th patent belongs to the j th technological class and 0 if not. When there is no room for confusion, we will drop the exponent part and write only p_{ij} when referring to either

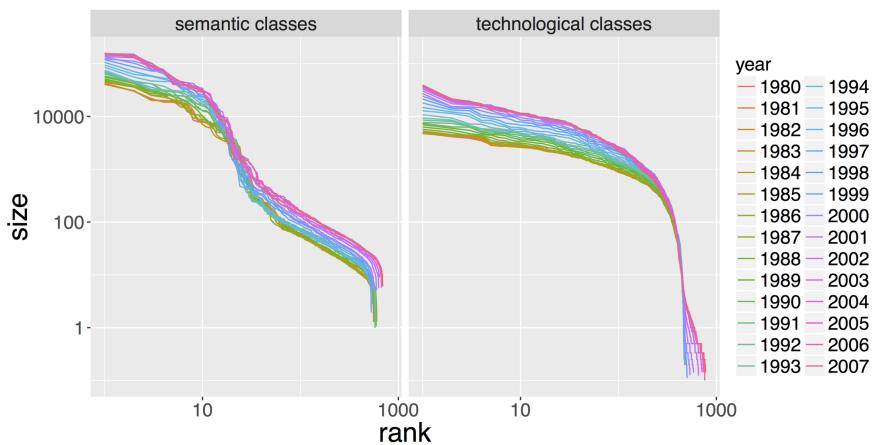


FIGURE 38 : Sizes of classes. Yearly from $t = 1980$ to $t = 2007$, we plot the size of semantic classes (left-side) and technological classes (right-side) for the corresponding time window $[t - 4, t]$, from the biggest to the smallest. The formal definition of size can be found in Section . Each color corresponds to one specific year. Yearly semantic classes and technological classes present a similar hierarchical structure which confirms the comparability of the two classifications. This feature is crucial for the statistical analysis in Section . Over time, curves are translated and levels of hierarchy stays roughly constant.

the technological or semantic matrix. Empirically, we find that both classes exhibit a similar hierarchical structure in the sense of a power-law type of distribution of class sizes as shown in Fig. 38. This feature is important, it suggests that a classification based on the text content of patents has some separating power in the sense that it does not divide up all the patents in one or two communities.

Potential Refinements of the Method

Our semantic classification method could be refined by combining it with other techniques such as Latent Dirichlet Allocation which is a widely used topic detection method (e.g. [BLEI, NG et JORDAN, 2003]), already used on patent data as in [KAPLAN et VAKILI, 2015] where it provides a measure of idea novelty and the counter-intuitive stylized facts that breakthrough invention are likely to come out of local search in a field rather than distant technological recombination. Using this approach should first help further evaluate the robustness of our qualitative conclusions (external validation). Also, depending on the level of orthogonality with our classification, it can potentially bring an additional feature to characterize patents, in the spirit of multi-modeling techniques where neighbor models are combined to take advantage of each point of view on a system.

Our use of network analysis can also be extended using newly developed techniques of hyper-network analysis. Indeed, patents and

keywords can for example be nodes of a bipartite network, or patents be links of an hyper-network, in the sense of multiple layers with different classification links and citation links. The combination of citation network modeling by Stochastic Block Modeling with topic modeling was studied for scientific papers by [ZHU et al., 2013b], outperforming previous link prediction algorithms. [IACOVACCI, WU et BIANCONI, 2015] provide a method to compare macroscopic structures of the different layers in a multilayer network that could be applied as a refinement of the overlap, modularity and statistical modeling studied in this paper. Furthermore, it has recently been shown that measures of multilayer network projections induce a significant loss of information compared to the generalized corresponding measure [DE DOMENICO et al., 2015], which confirms the relevance of such development that we left for further research.

An other potential research development would be to further exploit the temporal structure of our dataset. Indeed, large progress have recently been made in complex network analysis of time-series data (see [GAO, SMALL et KURTHS, 2017] for a review). For example, [GAO et al., 2015] develops a method to construct multiscale network from time series, which could in our case be a solution to identify structures in patents trajectories at different levels, and be an alternative to the single scale modularity analysis we use.

Results

In this section, we present some key features of our resulting semantic classification showing both complementary and differences with the technological classification. We first present several measures derived from this semantic classification at the patent level : Diversity, Originality, Generality (Section) and Overlapping (Section). We then show that the two classifications show highly different topological measures and strong statistical evidence that they feature a different model (Sections and).

Patent Level Measures

Given a classification system (technological or semantic classes), and the associated probabilities p_{ij} for each patent i to belong to class j (that were defined in Section), one can define a patent-level diversity measure as one minus the Herfindhal concentration index on p_{ij} by

$$D_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} p_{ij}^2, \text{ with } z \in \{\text{tec, sem}\}.$$

We show in Fig. 39 the distribution over time of semantic and technological diversity with the corresponding mean time-series. This is

carried with two different settings, namely including/not including patents with zero diversity (i.e. single class patents). We call other patents “complicated patents” in the following. First of all, the presence of mass in small probabilities for semantic but not technological diversity confirms that the semantic classification contains patent spread over a larger number of classes. More interestingly, a general decrease of diversity for complicated patents, both for semantic and technological classification systems, can be interpreted as an increase in invention specialization. This is a well-known stylized fact as documented in [ARCHIBUGI et PIANTA, 1992]. Furthermore, a qualitative regime shift on semantic classification occurs around 1996. This can be seen whether or not we include patents with zero diversity. The diversity of complicated patents stabilizes after a constant decrease, and the overall diversity begins to strongly decrease. This means that on the one hand the number of single class patents begins to increase and on the other hand complicated patents do not change in diversity. It can be interpreted as a change in the regime of specialization, the new regime being caused by more single-class patents.

More commonly used in the literature are the measures of originality and generality. These measures follow the same idea than the above-defined diversity in quantifying the diversity of classes (whether technological or semantic) associated with a patent. But instead of looking at the patent's classes, they consider the classes of the patents that are cited or citing. Formally, the originality O_i and the generality G_i of a patent i are defined as

$$O_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in I_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in I_i} p_{i'k}} \right)^2 \quad \text{and} \quad G_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in \tilde{I}_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in \tilde{I}_i} p_{i'k}} \right)^2,$$

where $z \in \{tec, sem\}$, I_i denotes the set of patents that are cited by the i th patent within a five year window (i.e. if the i th patent appears at year t , then we consider patents on $[t - T_0, t]$) when considering the originality and \tilde{I}_i the set of patents that cite patent i after less than five years (i.e. we consider patents on $[t, t + T_0]$) in the case of generality. Note that the measure of generality is forward looking in the sense that $G_i^{(z)}$ used information that will only be available 5 years after patent applications. Both measures are lower on average based on semantic classification than on technological classification. Fig. 40 plots the mean value of $O_i^{(sem)}$, $O_i^{(tec)}$, $G_i^{(sem)}$ and $G_i^{(tec)}$.

Classes overlaps

A proximity measure between two classes can be defined by their overlap in terms of patents. Such measures could for example be used

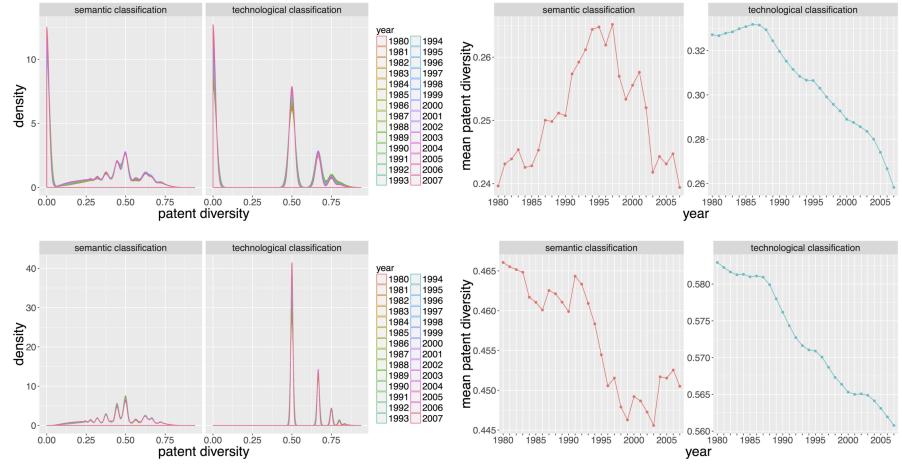


FIGURE 39 : Patent level diversities. Distributions of diversities (Left column) and corresponding mean time-series (Right column) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t - 4, t]$). The first row includes all classified patents, whereas the second row includes only patents with more than one class (i.e. patents with diversity greater than 0).

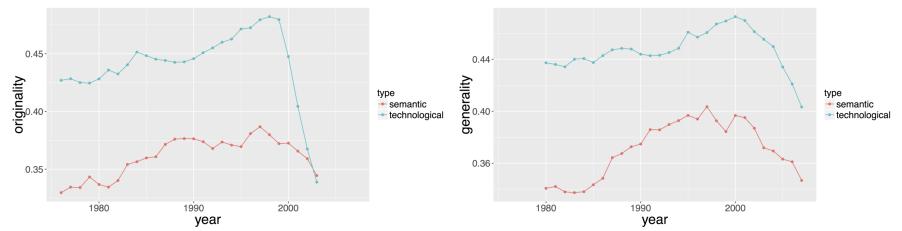


FIGURE 40 : Patent level originality (left hand side) and **generality** (right hand side) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t - 4, t]$) as defined in subsection .

to construct a metrics between semantic classes. Intuitively, highly overlapping classes are very close in terms of technological content and one can use them to measure distance between two firms in terms of technology as done in [BLOOM, SCHANKERMAN et REENEN, 2013]. Formally, recalling the definition of (p_{ij}) as the probability for the i th patent to belong to the j th class and N_P as the number of patents it writes

$$\text{Overlap}_{jk} = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} p_{ij} p_{ik}. \quad (21)$$

The overlap is normalized by patent count to account for the effect of corpus size : by convention, we assume the overlap to be maximal when there is only one class in the corpus. A corresponding relative overlap is computed as a set similarity measure in the number of patents common to two classes A and B, given by $\sigma(A, B) = 2 \cdot \frac{|A \cap B|}{|A| + |B|}$.

INTRACLASSIFICATION OVERLAPS The study of distributions of overlaps inside each classification, i.e. between technological classes and between semantic classes separately, reveals the structural difference between the two classification methods, suggesting their complementary nature. Their evolution in time can furthermore give insights into trends of specialization. We show in Fig. 41 distributions and mean time-series of overlaps for the two classifications. The technological classification globally always follow a decreasing trend, corresponding to more and more isolated classes, i.e. specialized inventions, confirming the stylized fact obtained in previous subsection. For semantic classes, the dynamic is somehow more intriguing and supports the story of a qualitative regime shift suggested before. Although globally decreasing as technological overlap, normalized (resp. relative) mean overlap exhibits a peak (clearer for normalized overlap) culminating in 1996 (resp. 1999). Looking at normalized overlaps, classification structure was somewhat stable until 1990, then strongly increased to peak in 1996 and then decrease at a similar pace up to now. Technologies began to share more and more until a breakpoint when increasing isolation became the rule again. An evolutionary perspective on technological innovation [ZIMAN, 2003] could shed light on possible interpretations of this regime shift : as species evolve, the fitness landscape first would have been locally favorable to cross-insemination, until each fitness reaches a threshold above which auto-specialization becomes the optimal path. It is very comparable to the establishment of an ecological niche [HOLLAND, 2012], the strong interdependency originating here during the mutual insemination resulting in a highly path-dependent final situation.

INTERCLASSIFICATION OVERLAPS Overlaps between classifications are defined as in (), but with j standing for the j th technological class

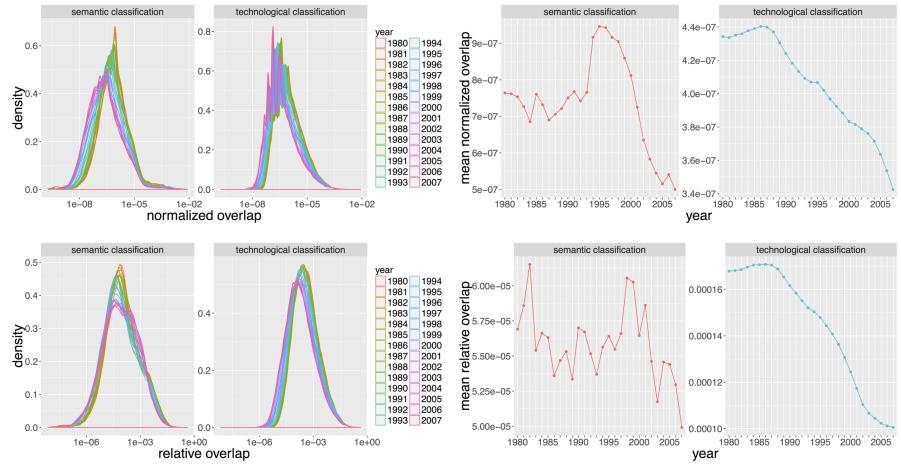


FIGURE 41 : **Intra-Classification overlaps.** (Left column) Distribution of overlaps O_{ij} for all $i \neq j$ (zero values are removed because of the log-scale). Right column) Corresponding mean time-series. (First row) Normalized overlaps. (Second row) Relative overlaps.

and k for the k th semantic class : p_{ij} are technological probabilities and p_{ik} semantic probabilities. They describe the relative correspondence between the two classifications and are a good indicator to spot relative changes, as shown in Fig. 42. Mean inter-classification overlap clearly exhibits two linear trends, the first one being constant from 1980 to 1996, followed by a constant decrease. Although difficult to interpret directly, this stylized fact clearly unveils a change in the *nature* of inventions, or at least in the relation between content of inventions and technological classification. As the tipping point is at the same time as the ones observed in the previous section and since the two statistics are different, it is unlikely that this is a mere coincidence. Thus, these observations could be markers of a hidden underlying structural changes in processes.

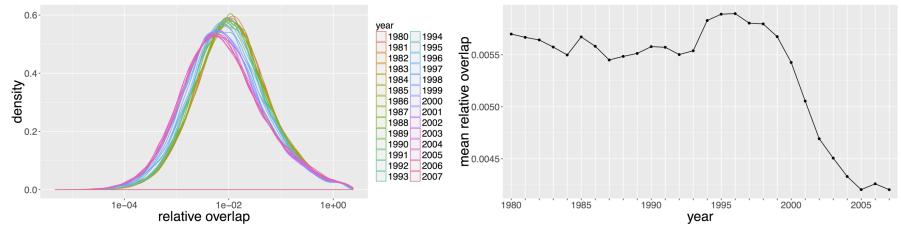


FIGURE 42 : **Distribution of relative overlaps between classifications.** (Left) Distribution of overlaps at all time steps; (Right) Corresponding mean time-series. The decreasing trend starting around 1996 confirms a qualitative regime shift in that period.

Citation Modularity

An exogenous source of information on relevance of classifications is the citation network described in Section ???. The correspondence between citation links and classes should provide a measure of accuracy of classifications, in the sense of an external validation since it is well-known that citation homophily is expected to be quite high (see, e.g, [ACEMOGLU et KERR, 2016]). This section studies empirically modularities of the citation network regarding the different classifications. To corroborate the obtained results, we propose to look at a more rigorous framework in Section . Modularity is a simple measure of how communities in a network are well clustered (see [CLAUSSET, NEWMAN et MOORE, 2004] for the accurate definition). Although initially designed for single-class classifications, this measure can be extended to the case where nodes can belong to several classes at the same time, in our case with different probabilities as introduced in [NICOSIA et al., 2009]. The simple directed modularity is given in our case by

$$Q_d^{(z)} = \frac{1}{N_p} \sum_{1 \leq i, j \leq N_p} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{N_p} \right] \delta(c_i, c_j),$$

with A_{ij} the citation adjacency matrix (i.e. $A_{ij} = 1$ if there is a citation from the i th patent to the j th patent, and $A_{ij} = 0$ if not), $k_i^{in} = |I_i|$ (resp. $k_i^{out} = |\tilde{I}_i|$) in-degree (resp. out-degree) of patents (i.e. the number of citations made by the i th patent to others and the number of citations received by the i th patent). Q_d can be defined for each of the two classification systems : $z \in \{\text{tec}, \text{sem}\}$. If $z = \text{tec}$, c_i is defined as the main patent class, which is taken as the first class whereas if $z = \text{sem}$, c_i is the class with the largest probability.

Multi-class modularity in turns is given by

$$Q_{ov}^{(z)} = \frac{1}{N_p} \sum_{c=1}^{N(z)} \sum_{1 \leq i, j \leq N_p} \left[F(p_{ic}, p_{jc}) A_{ij} - \frac{\beta_{i,c}^{out} k_i^{out} \beta_{j,c}^{in} k_j^{in}}{N_p} \right],$$

where

$$\beta_{i,c}^{out} = \frac{1}{N_p} \sum_j F(p_{ic}, p_{jc}) \text{ and } \beta_{j,c}^{in} = \frac{1}{N_p} \sum_i F(p_{ic}, p_{jc}).$$

We take $F(p_{ic}, p_{jc}) = p_{ic} \cdot p_{jc}$ as suggested in [NICOSIA et al., 2009]. Modularity is an aggregated measure of how the network deviates from a null model where links would be randomly made according to node degree. In other words it captures the propensity for links to be inside the classes. Overlapping modularity naturally extends simple modularity by taking into account the fact that nodes can belong simultaneously to many classes. We document in Fig. 43 both simple and multi-class modularities over time. For simple modularity,

$Q_d^{(tec)}$ is low and stable across the years whereas $Q_d^{(sem)}$ is slightly greater and increasing. These values are however low and suggest that single classes are not sufficient to capture citation homophily. Multi-class modularities tell a different story. First of all, both classification modularities have a clear increasing trend, meaning that they become more and more adequate with citation network. The specializations revealed by both patent level diversities and classes overlap is a candidate explanation for this growing modularities. Secondly, semantic modularity dominates technological modularity by an order of magnitude (e.g. 0.0094 for technological against 0.0853 for semantic in 2007) at each time. This discrepancy has a strong qualitative significance. Our semantic classification fits better the citation network when using multiple classes. As technologies can be seen as a combination of different components as shown by [YOUN et al., 2015], this heterogeneous nature is most likely better taken into account by our multi-class semantic classification.

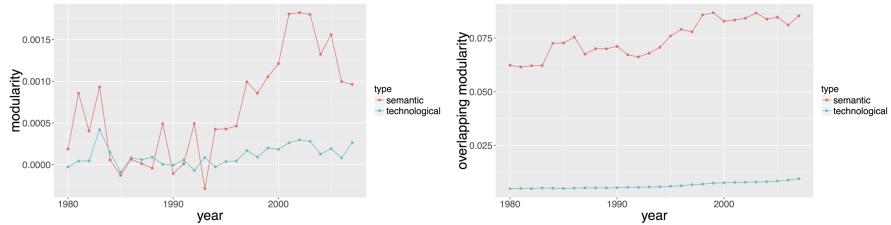


FIGURE 43 : **Temporal evolution of semantic and technological modularities of the citation network.** (Left) Simple directed modularity, computed with patent main classes (main technological class and semantic class with larger probability). (Right) Multi-class modularity, computed following [NICOSIA et al., 2009]

Statistical Model

In this section, we develop a statistical model aimed at quantifying performance of both technological and semantic classification systems. In particular, we aim at corroborating findings obtained in Section . The mere difference between this approach and the citation modularity approach lies in the choice of the underlying model, and the according quantities of interest. In addition for the semantic approach, we want to see if when restricting to patents with higher probabilities to belong to a class, we obtain better results. To do that, we choose to look at within class citations proportion (for both technological and semantic approaches). We provide two obvious reasons why we choose this. First, the citations are commonly used as a proxy for performance as mentioned in Section . Second, this choice is “statistically fair” in the sense that both approaches have focused on various

goals and not on maximizing directly the within class proportion. Nonetheless, the within class proportion is too sensitive to the distribution of the shape of classes. For example, a dataset where patents for each class account for 10% of the total number of patents will mechanically have a better within class proportion than if each class accounts for only 1%. Consequently, an adequate statistical model, which treats datasets fairly regardless of their distribution in classes, is needed. This effort ressembles to the previous study of citation modularity, but is complementary since the model presented here can be understood as an elementary model of citation network growth. Furthermore, the parameters fitted here can have a direct interpretation as a citation probability.

We need to introduce and recall some notations. We consider a specific window of observations $[t - T_0, t]$, and we define Z the number of patents which appeared during that time window. We let t_1, \dots, t_Z their corresponding appearance date by chronological order, which for simplicity are assumed to be such that $t_1 < \dots < t_Z$. For each patent $i = 1, \dots, Z$ we consider C_i the number of distinctive couples {cited patent, cited patent's class} made by the i th patent (for instance if the i th patent has only made one citation and that the cited patent is associated with three classes, then $C_i = 3$). Let $z \in \{\text{tec}, \text{sem}\}$, we define $N_i^{(z)}$ the number of patents associated to at least one of the i th classes at time t_{i-1} . For $l = 1, \dots, C_i$ we consider the variables $B_{l,i}$, which equal 1 if the cited patent's class is also common to the i th patent. We assume that $B_{l,i}$ are independent of each other and conditioned on the past follow Bernoulli variables

$$B\left(\min\left\{1, \frac{N_i^{(z)}}{i-1} + \theta^{(z)}\right\}\right),$$

where the parameter $0 \leq \theta^{(z)} \leq 1$ indicates the propensity for any patent to cite patents of its own technological or semantic class. When $\theta^{(z)} = 0$, the probability of citing patents from its own class is simply $N_i^{(z)}(i-1)^{-1}$, which corresponds to the observed proportion of patents which belong to at least one of the i th patent's classes. Thus this corresponds to the estimated probability of citing one patent if we assume that the probability of citing any patent $k = 1, \dots, i-1$ is uniformly distributed, which could be a reasonable assumption if classes were assigned randomly and independently from patent abstract contents. Conversely if $\theta^{(z)} = 1$, we are in the case of a model where there are 100% of within class citations. A reasonable choice of $\theta^{(z)}$ lies between those two extreme values. Finally, we assume that the number of distinctive couples C_i are a sequence of independent and identically distributed random variables following the discrete distribution C , and also independent from the other quantities.

We estimate $\theta^{(z)}$ via maximum likelihood, and obtain the corresponding maximum likelihood estimator (MLE) $\hat{\theta}^{(z)}$. The likelihood

function, along with the standard deviation expression and details about the test, can be found in ???. The fitted values, standard errors and p-values corresponding to the statistical test $\theta^{(sem)} = \theta^{(tec)}$ (with corresponding alternative hypothesis $\theta^{(sem)} > \theta^{(tec)}$) on non-overlapping blocks from the period 1980-2007 are reported on Table 6. Note that the estimation included patents up until 2010 in the period 2006-2007 and not the patents from 1980 in the period 1980-1985 for homogeneity in size with other periods. This doesn't affect the significance of the results. Semantic values are reported for four different chosen thresholds $p^- = .04, .06, .08, .1$. It means that we restricted to the couples (ith patent, jth class) such that $p_{ij} \geq p^-$.

The choice of considering non-overlapping blocks (instead of overlapping blocks) is merely statistical. Ultimately, our interest is in the significance of the test over the whole period 1980-2007. Thus, we want to compute a global p-value. This can be done considering the local p-values (by local, we mean for instance computed on the period 2001-2005) assuming independence between them. This assumption is reasonable only if the blocks are non-overlapping. All of this can be found in ???. Finally, note that from a statistical perspective, including overlapping blocks wouldn't yield more information.

The values reported in Table 6 are overwhelmingly against the null hypothesis. The global estimates of $\theta^{(sem)}$ are significantly bigger than the estimate of $\theta^{(tec)}$ for all the considered thresholds. Although the corresponding p-values (which are also very close to 0) are not reported, it is also quite clear that the bigger the threshold, the higher the corresponding $\theta^{(sem)}$ is estimated. This is consistently seen for any period, and significant for the global period. This seems to indicate that when restricting to the couples (patent, class) with high semantic probability, the propension to cite patents from its own class $\theta^{(sem)}$ is increasing. We believe that this might provide extra information to patent officers when making their choice of citations. Indeed, they could look first to patents which belong to the same semantic class, especially when patents have high probability semantic values.

Note that the introduced model can be seen as a simple model of citations network growth conditional to a classification, which can be expressed as a stochastic block model (e.g. [DECCELLE et al., 2011], [VALLES-CATALA et al., 2016]). The parameters are estimated computing the corresponding MLE. In view of [NEWMAN, 2016], this can be thought as equivalent to maximizing modularity measures.

Conclusion

The main contribution of this study was twofold. First we have defined how we built a network of patents based on a classification that uses semantic information from abstracts. We have shown that this classification share some similarities with the traditional techno-

logical classification, but also have distinct features. Second, we provide researchers with materials resulting from our analysis, which includes : (i) a database linking each patent with its set of semantic classes and the associated probabilities ; (ii) a list of these semantic classes with a description based on the most relevant keywords ; (iii) a list of patent with their topological properties in the semantic network (centrality, frequency, degree, etc.). The availability of this data suggests new avenues for further research. Linking our dataset with existing open ones can lead to various powerful developments. For example, using it together with the disambiguated inventor database provided by [LI et al., 2014] could be a way to study semantic profiles of inventors, or of cities as inventor addresses are provided. The investigation of spatial diffusion of innovation between cities, which is a key component of Pumain's Evolutive Urban Theory [PUMAIN, 2010], would be made possible.

A first potential application is to use the patents' topological measures inherited from their relevant keywords. The fact that these measures are backward-looking and immediately available after the publication of the patent information is an important asset. It would for example be very interesting to test their predicting power to assess the quality of an innovation, using the number of forward citations received by a patent, and subsequently the future effect on the firm's market value.

Regarding firm innovative strategy, a second extension could be to study trajectories of firms in the two networks : technological and semantic. Merging these information with data on the market value of firms can give a lot of insight about the more efficient innovative strategies, about the importance of technology convergence or about acquisition of small innovative firms. It will also allow to observe innovation pattern over a firm life cycle and how this differ across technology field.

A third extension would be to use dig further into the history of innovation. USPTO patent data have been digitized from the first patent in July 1790. However, not all of them contain a text that is directly exploitable. We consider that the quality of patent's images is good enough to rely on Optical Character Recognition techniques to retrieve plain text from at least 1920. With such data, we would be able to extend our analysis further back in time and to study how technological progress occurs and combines in time. [AKCIGIT, KERR et NICHOLAS, 2013] conduct a similar work by looking at recombination and apparition of technological subclasses. Using the fact that communities are constructed yearly, one can construct a measure of proximity between two successive classes. This could give clear view on how technologies converged over the year and when others became obsolete and replaced by new methods.

TABLE 6 : Estimated values of $\theta^{(tec)}$ and $\theta^{(sem)}$ and corresponding standard errors obtained from a Maximum Likelihood estimator as presented in section .

Approach	Estimated Value	st. er.	p-value
1980-1985 period			
technological	.664	.008	
semantic $p^- = .04$.741	.047	.053
semantic $p^- = .06$.799	.081	.049
semantic $p^- = .08$.828	.126	.097
semantic $p^- = .10$.834	.166	.153
1986-1990 period			
technological	.634	.007	
semantic $p^- = .04$.703	.022	.001
semantic $p^- = .06$.768	.040	.0004
semantic $p^- = .08$.804	.069	.007
semantic $p^- = .10$.832	.114	.041
1991-1995 period			
technological	.619	.006	
semantic $p^- = .04$.655	.009	.0004
semantic $p^- = .06$.713	.017	9e-08
semantic $p^- = .08$.731	.025	7e-06
semantic $p^- = .10$.750	.037	9e-06
1996-2000 period			
technological	.551	.003	
semantic $p^- = .04$.585	.002	≈ 0
semantic $p^- = .06$.638	.004	≈ 0
semantic $p^- = .08$.660	.006	≈ 0
semantic $p^- = .10$.686	.008	≈ 0
2001-2005 period			
technological	.567	.003	
semantic $p^- = .04$.621	.004	≈ 0
semantic $p^- = .06$.676	.007	≈ 0
semantic $p^- = .08$.701	.010	≈ 0
semantic $p^- = .10$.710	.013	≈ 0
2006-2007 period			
technological	.600	.007	
semantic $p^- = .04$.683	.016	1e-06
semantic $p^- = .06$.732	.025	2e-07
semantic $p^- = .08$.760	.036	6e-06
semantic $p^- = .10$.782	.048	9e-05
1980-2007 global period			
technological	.606	.002	
[semantic $p^- = .04$ - Thesis version 3.1]	.665	.009	8e-11
semantic $p^- = .06$.721	.017	9e-12
semantic $p^- = .08$.747	.025	9e-09

D

DONNÉES

This appendix lists and describes the different open datasets created and used in the thesis.

TODO : when possible, specify data citation (ex. traffic data : TransportationEquilibrium paper); try to put all on dataverse ; laius sur dataverse, partage des données etc.

Les données comme domaine de connaissance propre : décrire opération de collecte des données et de construction des jeux.

D.1 DONNÉES DE TRAFFIC DU GRAND PARIS

D.2 PRIX DE L'ESSENCE AUX ETATS-UNIS

D.3 RÉSEAU ROUTIER EUROPÉEN

D.4 RÉSEAU DYNAMIQUE DES AUTOROUTES FRANÇAISES

TODO : Merger avec la base bassin parisien de Florent, faire un data paper.

D.5 INTERVIEWS

TODO : Possible interview in Guandong : Zhuhai Planning Bureau (people at the workshop); Hong Kong Transportation authority (see demand in name of Medium : how to proceed?) : easy for english ?

E

OUTILS

E.1 SOFTWARES AND PACKAGES

This appendix lists and describes the different open datasets created and used in the thesis.

E.1.1 *largeNetwoRk : Import de réseau et simplification pour R*

E.1.2 *Fouille de Corpus scientifique*

E.2 ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION

You must not be afraid of putting code in your thesis, code is not dirty
 - ALEXIS DROGOUL PhD defense
 of [REY-COYREHOURCQ, 2015]

And yet it is. It makes no sense to put code listings in the core of the text if there is no particular algorithmic detail that requires attention. As soon as implementation biases are avoided, architecture and source for a computational model should be independent from its formal description (but provided along model description with source code as already mentioned before). We give in this appendix architectural details on main models of simulation or algorithms we used. Langage and size (in code lines) are provided, along with architectural remarkable features. See <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models> for all models, empirical analysis and small experiments. The following reports are partially generated automatically using experimental tools aimed at workflow improvement.

E.2.1 Revue Systématique Algorithmique

OBJECTIFS Implement systematic literature review algorithm.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp>

CARACTÉRISTIQUES

- Language : Java
- Size : 7116

PARTICULARITÉS

- HashConsing used for unique bibliography object, specific hashCode switching if id available or only titles (proceed to lexical distance comparison in that latest case).
- API to context currently being replaced by Python scripts.

ARCHITECTURE Classical object oriented, see code.

SCRIPTS ADDITIONNELS R for result exploration and visualization.

E.2.2 Bibliométrie Indirecte

OBJECTIFS Hypernetworks analysis of cybergeo journal.

LOCALISATION <https://github.com/Geographie-cites/cybergeo20/tree/master/HyperNetwork>
<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp> for common Java part.

CARACTÉRISTIQUES

- Language : Python, R and Java.
- Size : -

PARTICULARITÉS Polyglot

ARCHITECTURE See schema chapter 3.

SCRIPTS ADDITIONNELS -

E.2.3 Croissance Urbaine

OBJECTIF Simple density urban growth model.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Density>

CARACTÉRISTIQUES

- Language : NetLogo then scala.
- Size : 4355

PARTICULARITÉS Morphological indicators in scala implemented with Fast Fourier transform ; with R communication in NetLogo.

ARCHITECTURE Nothing particular.

SCRIPTS ADDITIONNELS R for result exploration and morphological analysis.
 oms for model exploration.

E.2.4 Génération des Données Synthétiques Corrélées

OBJECTIFS Weak coupling of density generation and network generation.

LOCALISATION https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Network_20151229

CARACTÉRISTIQUES

- Language : NetLogo (network) and scala.
- Size : 3188

PARTICULARITÉS Network heuristic easier to implement and explore in netlogo

ARCHITECTURE OpenMole allows coupling between modules through exploration script.

SCRIPTS ADDITIONNELS R for result exploration.
oms for model exploration.

E.2.5 Modèle Lutecia

OBJECTIF Implementation of Lutecia model, chapter ??.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Governance/MetropolSim/Lutecia>

CARACTÉRISTIQUES

- Language : NetLogo
- Size : 4791

PARTICULARITÉS Shortest path dynamical programming using matrices.

ARCHITECTURE Pseudo object architecture in agent environment.

SCRIPTS ADDITIONNELS R for result exploration.
oms for model exploration.

E.2.6 Analyse des Réseaux

OBJECTIF Simplification of european road network

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/StaticCorrelations>

CARACTÉRISTIQUES

- Language : R, Shell, PostgreSQL
- Size : 505

PARTICULARITÉS Handling of large size databases imposes sequential processing ; use of external program osmosis for conversion from osm data to pgsql.

ARCHITECTURE Shell script lead maneuvers.

SCRIPTS ADDITIONNELS -

E.3 TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH

Open for Discovery
- PLoS

We briefly evoke here tools or workflows currently under development or testing, aimed at easing an open reproducible research and making it more transparent.

E.3.1 *Générateur de Documentation Netlogo*

Documentation generation is central for reproducibility as it can automatize implementation description. NetLogo does not provide a documentation generator and we are thus currently writing a Doxygen wrapper for NetLogo code, that basically consists in transforming NetLogo code into Java code and parsing documentation comment blocks. An experimental version is available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Doc>.

E.3.2 *git comme outil de reproductibilité*

The use if git as a reproducibility and transparency tool was emphasized in [RAM, 2013] (for various reasons such as exact history tracing, easy cloning, past commit branching). It furthermore can help individual workflow for advantages such as automatic backup, organisation, experiments tracking. We use it actively and develop extensions for it.

E.3.3 *git-data*

git-data is a shell based (experimental) git extension, available at <https://github.com/JusteRaimbault/gitdata>, that allows automated backup of large file within a git repository, their transparent integration in ignored files and the creation of symbolic links for a transparent local use.

E.3.4 *Vers un gestionnaire de métadonnées compatible avec git*

The issue of meta-data for figures is a crucial issue, as it is often difficult to keep a trace of all parameter values that have generated it, along with the corresponding code. Tricks may furthermore happen in script environments such as R or python when variables are

accidentally modified without code modification. Keeping an exhaustive trace of the exact dataset, code and history that has generated a precise figure is a necessary condition for exact reproducibility. We are elaborating a git-compatible tool that would automatically handle these metadata, for example by branching and associating the unique commit hash to the figure. To become not an organizational burden nor a repository perturbation, we must still make some experiments. The final idea would be to have under each figure a unique identifier linking to the associated reproducing environment.

E.3.5 *TorPool*

TorPool is a java based Tor wrapper available with an api (currently only java, R version projected) at <https://github.com/JusteRaimbault/TorPool>. It allows among other purposes tricky data retrieval.



QUANTITATIVE ANALYSIS OF THESIS
REFLEXIVITY

Quantitative Analysis of Thesis reflexivity

TODO : faire un graphe des concepts; compare to semantic network of concepts in Gödel Escher Bach.